Connecting variation in genome structure and chromatin composition in Zea *mays*

A DISSERTATION SUBMITTED
TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Jaclyn M Noshay

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

Dr. Nathan M Springer

February 2021

Acknowledgements

I would like to thank my advisor, Nathan Springer for his support and encouragement during my time in his lab. I have grown academically and personally throughout my time as a PhD student and my ability to explore the various avenues of my research and potential career paths was greatly appreciated. I would also like to thank the members of my advisory committee including Jane Glazebrook, Peter Tiffin, and Cory Hirsch for input on my dissertation research and support throughout the process. I received tremendous support from faculty and staff of the Plant and Microbial Biology Program as well as additional members of the Springer lab. I would like to personally thank Pete Hermanson, Sara Eliason, Sarah Anderson, Pete Crisp and Candy Hirsch for their support through the various challenges of graduate school.

My network of friends beyond academics was an essential piece of my time at the University of Minnesota. I would like to thank my best friend Kayla Masessa for listening, supporting, and putting up with me. I would also like to thank my Mill City Running community for providing me with an outlet and escape from the ups and downs of my PhD through endless miles and smiles.

Lastly, I would like to thank my family. My mom and dad for teaching me the power of education and the love for questioning at a young age, for always supporting me through the exciting times and the rough times, for encouraging me to find my passion and making sure I got back up every time I was knocked down. I owe much of my success and sanity to the love and encouragement of my parents. My brothers, Ryan and Michael, for being incredible role models, challenging me to find my path in life, and helping me see the positives along the way.

Introduction

In many crop species there is tremendous intraspecific variation for genome structure due to highly variable transposon insertions. The goal of my thesis research is to provide insight on genetic and epigenetic dynamics and their relationship with functional variation which has the potential to influence variation in regulation and gene expression. 'Epigenetics' describes heritable information not solely due to the DNA sequence, whereas 'genetics' is heritable information directly related to the DNA sequence. The central question of the chapters presented is to ask, what are the relative contributions of genetic (transposable element insertions) and epigenetic (localized chromatin changes) factors to variation in DNA methylation and gene expression?

In order to address this topic, I first present background information on both DNA methylation and epigenetic influence in maize. It is pertinent to understand the many roles and mechanisms of DNA methylation in plant species in order to decipher the contributions to variation. While this DNA methylation has previously been assessed, the ability to tease apart epigenetics from genetics through polymorphism detection on a genome-wide scale is possible only with new technology. These advancements have helped us to understand information found within the maize epigenome which may not be captured by genetic variation and therefore provide additional data for predicting traits and improving the efficiency of plant improvement strategies.

I have conducted several research studies to address the question of epigenetic stability in maize. To first address the dynamic between DNA methylation and transposable elements across the genome, I sought to characterize epigenetic patterns associated with TE families and the cause or effect of TE insertion on DNA methylation architecture.
Chapter III presents the assessment of natural variation of transposon insertions and the impact on epiallele state.

After identifying how these TEs interact with their flanking sequence I further questioned the genomic influence of the TE body in chapter IV. Accessible chromatin data has allowed identification of putative regulatory regions genome-wide and I pursued the question of how novel TE sequence can contribute to the regulatory dynamics of an organism. Through polymorphic TE insertions we were able to assess the influence of these enhancers on nearby gene expression.

The final chapter of my thesis seeks to question the stability of the maize methylome. Now focusing on the shared and nonshared sequence between maize genotypes, I was able to analyze epigenetic variation in the presence or absence of sequence variation. A pan-genome study allows for identification of both core (present across all genotypes) and dispensable (variable between

genotypes) epigenetic regions. Presence of variable methylation state is indicative of epigenetic patterns not predictable by sequence.

The work presented below describes these broad inquiries in further detail working to answer essential questions regarding genetic and epigenetic contributions to maize phenotype

Table of Contents

List of Tables

List of Figures

CHAPTER V:

CHAPTER I: Context Statement

DNA methylation is a chromatin modification that has generally been associated with gene silencing or heterochromatin. Plants have mechanisms to allow for the stable inheritance of DNA methylation through mitosis or meiosis. This creates the potential for DNA methylation to provide epigenetic inheritance for traits in maize and other crops. Epigenetics refers to heritable transmission of information that is not solely attributable to DNA sequence. Several examples of epigenetic inheritance were first described in maize including paramutation, imprinting, and transposable element inactivation. There is evidence that DNA methylation is associated with each of these epigenetic phenomena.

In addition, natural variation for epigenetic states may contribute substantially to variation among maize inbred lines and could be an important source of variation for crop improvement. Advances in our understanding of the molecular mechanisms controlling DNA methylation in Arabidopsis have provided clues to the genes and pathways likely to be important in maize. Recent technological developments have provided the opportunity to characterize the genome-wide distribution of DNA methylation in the maize genome. This has provided insights into the patterns of DNA methylation in plant species with large, complex genomes and has led to the identification of potential cryptic genomic information that is silenced by DNA methylation. We will summarize current understanding of the mechanisms that regulate methylation and factors that influence variation and stability of the maize methylome.

Chapter I entitled ‘Literature review of the maize methylome’ has been adapted from my work in the publication:

Jaclyn Noshay*, Peter Crisp, Nathan Springer (2018).* The Maize Methylome. *The Zea Mays Genome.*

During the course of this work several authors have made contributions. Jaclyn M Noshay, Peter A Crisp, and Nathan M Springer did extensive background research, discussed and executed writing, and generated summary analyses and figures. I have removed author contact information and acknowledgments as well as reformatted figures and references to be consistent throughout my thesis.

CHAPTER I:

Literature review of the maize methylome

Introduction

In maize, as in other eukaryotes, DNA methylation refers to the addition of a methyl group to the 5' carbon of cytosine residues. This methyl group is added after DNA replication. Therefore, the faithful maintenance of DNA methylation patterns requires mechanisms to copy DNA methylation onto the daughter strand. A large majority of DNA methylation in maize, and other plants, is found at CG or CHG (where H is any base except G) sites that have symmetry across the two strands of DNA (Niederhuth et al. 2016). This allows for the maintenance of DNA methylation through targeted methylation of hemi-methylated DNA that results from the incorporation of unmethylated cytosines during DNA replication. Cytosine residues that are not followed by a G in the next two bases (CHH sites) can also be methylated but require alternative mechanisms for maintenance of the patterns following replication (Law and Jacobsen 2010; Matzke and Mosher 2014; Springer and Schmitz 2017). In recent years, Arabidopsis has provided a model system for studying DNA methylation due to the availability of reverse genetics resources and the viability of mutants with severely reduced DNA methylation (Law and Jacobsen 2010; Matzke and Mosher 2014). Our knowledge of the specific mechanisms that control DNA methylation and the role of DNA methylation in maize and other crop plants is more limited. Here we will describe what is known in maize and contrast with data from Arabidopsis noting both conserved features and key differences.

Methods for documenting DNA methylation

There are a variety of approaches that have been utilized to monitor DNA methylation, with varying levels of sensitivity and specificity (reviewed by Zilberman et al. 2007). The genomic proportion of cytosine residues that are methylated can be roughly estimated by HPLC (Papa 2001). This approach is useful for quantifying genome-wide DNA methylation levels but it cannot

determine the level of methylation at specific sequence contexts, sites, or regions in the genome. In many cases the presence of DNA methylation can inhibit digestion by restriction enzymes and in some rare cases there are restriction enzymes (McrBC, FspEI, MspJI) that require DNA methylation in order to cut a site (Loenen and Raleigh 2014). These methylation-sensitive or –dependent enzymes can be combined with Southern blotting or quantitative PCR approaches to document the presence or absence of methylation at specific sites in the genome (Zhang et al. 2014). In general, the use of restriction enzymes for surveying DNA methylation can provide data for specific sites but tends to be only partially quantitative and can be difficult to apply in a high-throughput fashion. Methylation-sensitive enzymes can be combined with AFLP- based approaches to provide a survey of methylation at many different sites (Lu et al. 2008). Methylation-dependent enzymes have been used in combination with shotgun sequencing or microarray approaches for genome-wide identification of unmethylated regions referred to as methylation filtration (Palmer 2003; Rabinowicz et al. 2005). Another approach for documenting genome-wide methylation levels utilizes a 5-methylcytosine antibody for immunoprecipitation of methylated DNA (meDIP) (Eichten et al. 2011). This approach enriches for fragments containing DNA methylation and can be combined with microarrays or high-throughput sequencing approaches to provide genome-wide profiles. The methylation filtration and meDIP assess regional methylation throughout a genome but do not provide single-base resolution of DNA methylation.

The “gold-standard” approach for measuring DNA methylation is with sodium bisulfite treatment followed by sequencing (Lister et al. 2008). Treatment of single-stranded DNA with sodium bisulfite will result in conversion of unmethylated cytosine residues to uracil while methylated cytosines are not converted. Sequencing of treated molecules reveals which bases remained as cytosine (methylated) and which bases were converted (unmethylated). By sequencing multiple molecules, the frequency of methylation at any particular

site can be determined. This approach was initially combined with PCR to document methylation at particular genomic regions. In recent years, this has been paired with next-generation sequencing to perform whole genome bisulfite sequencing (WGBS) which provides base-level resolution and context-specific information for DNA methylation throughout the portion of the genome for which unique alignments are possible (Lister et al. 2008; Regulski et al. 2013; Gent et al. 2013). Bisulfite treatment can also be paired with sequence capture approaches to provide single-base resolution for a subset of genomic regions (Lit et al. 2015c).

Genomic distribution of DNA methylation in maize

WGBS has been used to document the genome-wide distribution of DNA methylation in maize (Regulski et al. 2013; Gent et al. 2013; West et al. 2014). However, it is worth noting that current short-read sequencing and bioinformatics approaches cannot interrogate the entire genome. WGBS allows analysis of regions covered by uniquely aligning reads, which results in coverage for ~70% of the maize genome. Genic (78% coverage) and intergenic (90% coverage) regions have substantially higher coverage than TEs (60% coverage) for methylation data (Figure 1). WGBS profiles have revealed that plant genomes have similar mechanisms for DNA methylation but the frequency and patterning of methylation domains varies among species (Niederhuth et al. 2016; Springer and Schmitz 2017). While maize has most of the methylation machinery found in Arabidopsis it must operate to methylate a genome with a different organization. Arabidopsis has a relatively small genome with a high gene density (The Arabidopsis Genome Initiative 2000), most genes are not located near TEs, and the vast majority of heterochromatin in the Arabidopsis genome is located in pericentromeric regions. In contrast, the maize genome has a much lower gene density (Schnable et al. 2009; Jiao et al. 2017) and TEs are prevalent throughout the whole length of maize chromosomes (Baucom et al. 2009). The total abundance and relative distribution of CG, CHG, and CHH across the genomes of Arabidopsis and

maize are distinct (West et al. 2014; Neiderhuth et al. 2016). Methylation in all three sequence contexts is highly enriched within TEs, repeat sequences, and pericentromeric regions in Arabidopsis (West et al. 2014). Maize has among the highest levels of CG and CHG methylation in species with characterized methylation profiles, and methylation at CG and CHG contexts are found throughout the length of the maize chromosome (Springer and Schmitz 2017; West et al. 2014; Nierderhuth et al. 2016). In contrast, the levels of CHH methylation in maize are relatively low compared to many other plant species (West et al. 2014; Niederhuth et al. 2016).

***Methylation domains in the maize genome**:*

Assessing the relative levels of CG, CHG, and CHH methylation in windows of the maize genome can be used to define different types of methylation domains (Springer and Schmitz 2017). The methylation domains include CG/CHG/CHH regions (RNA directed DNA Methylation, or RdDM targets), CG/CHG regions (heterochromatin), CG only (gene body methylation – gBM), unmethylated regions, and unclassified regions with intermediate levels of DNA methylation (Figure 1). CG/CHG domains, which contain high levels of CG and CHG methylation, but very low levels of CHH methylation, are the most common type in the maize genome, accounting for large portions of intergenic and TE regions of the genome but are less abundant within genes (Figure 1). The RdDM targets, which have elevated methylation in all three contexts, only account for 2% of the maize genome and are most prevalent within intergenic regions. Regions with only CG methylation account for ~6% of the maize genome and are often found within maize gene bodies. Approximately 11% of the maize genome has low levels of methylation in all three contexts and this is most prevalent within the genic portions of the maize genome and is quite rare in TEs. Another 10% of the maize genome has intermediate levels of DNA methylation that are difficult to classify.

DNA methylation patterns at maize genes:

The distribution of methylation within plant genomes reflects the distinct methylation profiles at genes and TEs. In general, CG and CHG methylation levels are high in non- genic regions but drop to low levels near the transcription start site (TSS) and transcription termination site (TTS) of annotated genes (Regulski et al. 2013; Gent et al. 2013; West et al. 2014). Within gene bodies there is moderate levels of CG methylation likely reflecting gene body methylation (Neiderhuth et al. 2016). Maize also contains significant levels of CHG methylation in gene bodies that is partially attributable to methylation of TEs found within introns (West et al. 2014). CHH methylation is enriched in regions flanking maize genes (Gent et al. 2013). These mCHH islands mark the boundary between high levels of CG and CHG methylation outside of maize genes and the reduced levels of methylation in genes (Li et al. 2015a). Several factors influence the profile of DNA methylation over maize genes. In general, highly expressed genes have the lowest levels of DNA methylation at the TSS and TTS (Regulski et al. 2013; Gent et al. 2013; West et al. 2014). However, the inverse pattern is observed for CHH methylation in regions upstream of the TSS (Gent et al. 2013). Genes located in syntenic positions relative to other grasses exhibit much lower levels of DNA methylation than inserted (non-syntenic) genes (Eichten et al. 2011; West et al. 2014). There is no evidence for differential levels of DNA methylation for genes in the two sub-genomes that have resulted from the ancient whole genome duplication event in maize (Eichten et al. 2011; West et al. 2014).

DNA methylation patterns at maize TEs*:*

DNA methylation at TEs is high relative to flanking regions (West et al. 2014). The levels of CG and CHG methylation over TEs is higher in maize than in Arabidopsis (West et al. 2014), with more gradual transitions from low to high methylation levels at the edges of TEs, suggesting greater spreading of DNA methylation from TEs to flanking regions in maize (Eichten et al. 2012). The

analysis of transposon superfamilies revealed variation in chromatin profiles (West et al. 2014). While CG and CHG methylation are very high for all families there is variation for the level of CHH methylation and H3K9me2 (West et al. 2014). There is also evidence for family specific variation in whether DNA methylation can spread to flanking regions, suggesting that TE families associated with spreading are more likely to reduce the expression of nearby genes than families without spreading (Eichten et al. 2012). The association of CG and CHG methylation (inactive transcription) with spreading retrotransposon families and CHH (active transcription) with non-spreading retrotransposon families can explain this gene expression correlation. The methylation levels of transposons located within maize genes are quite similar to the levels for intergenic TEs even though these regions undergo active transcription (West et al. 2014). This suggests that methylated TEs do not pose a barrier to transcriptional elongation. However, there is evidence that plants require machinery to allow for proper transcription and splicing of regions that are highly methylated (To et al. 2015).

Molecular mechanisms regulating DNA methylation

DNA methylation at any locus is influenced by a variety of processes including methylation maintenance, de novo methylation, and demethylation. We will describe the mechanisms expected to control CG, CHG and CHH methylation based on studies in Arabidopsis and the evidence for similar systems being present in maize. The Arabidopsis genome encodes seven DNA methyltransferases including *DOMAINS REARRANGED METHYLASE 2* (*DRM1*) and *DRM2*, *CHROMOMETHYLASE 1* (*CMT1*), *CMT2*, and *CMT3*, *METHYLTRANSFERASE 1* (*MET1*) and *MET2*. Four of these methyltransferases (*DRM2*, *CMT2*, *CMT3*, and *MET1*) are responsible for the bulk of methylation in Arabidopsis and contribute to different maintenance and *de novo* methylation pathways (Stroud et al. 2013; Law and Jacobsen 2010, Matzke and Mosher 2014, Du et al. 2015).

CG methylation:

Genetic analysis has shown that *MET1* is required for CG methylation maintenance in Arabidopsis (Law and Jacobsen 2010). MET1 is dependent on three VARIANTS IN METHYLATION (VIM) proteins, which are ubiquitin E3 ligases containing an SRA domain that binds hemi-methylated DNA (Du et al. 2015; Woo et al. 2008). After MET1 is recruited to hemi-methylated CG sites, it functions to methylate the opposing strand, providing a robust mechanism to transmit CG methylation patterns following DNA replication. In maize, two tandem duplicates of MET1-like genes (*Zmet1* - Zm00001d018976 and Zm00001d018977) have been identified (Li et al. 2014a). The maize genome also encodes at least three VIM1-like genes. The tandemly duplicated MET1-like genes in maize likely play critical roles in maintaining CG methylation in the maize genome similar to *MET1* in Arabidopsis. To date, loss of function alleles have not been isolated for these genes through forward or reverse genetics approaches, limiting functional studies of these genes in maize.

CHG methylation:

In Arabidopsis, the bulk of CHG methylation is maintained by the chromomethylase *CMT3* (Matzke and Mosher 2014; Du et al. 2015; Bewick et al. 2016). CMT3 contains a BAH domain, a DNA methyltransferase domain, and a chromodomain. The chromodomain and BAH domain provide the ability for CMT3 to bind to histone H3 that has dimethylated lysine 9 (H3K9me2) (Du et al. 2012). In Arabidopsis, the enzyme that provides H3K9me2, KRYPTONITE (KYP), binds to CHG methylation (Du et al. 2014). This provides a self-reinforcing loop between CHG DNA methylation and H3K9me2 which provides a mechanism for stable memory of this chromatin state (Du et al. 2015). The maize genome encodes two paralogs that are related to Arabidopsis *CMT3*; *Zmet2* (*Dmt102* - Zm00001d026291) and *Zmet5* (*Dmt105* - Zm00001d002330) (Papa 2001; Makarevitch et al. 2007). A loss-of-function allele, *zmet2-m1*, results in significant reductions of genomic CHG

methylation levels (Papa 2001). Other partial loss-of- function alleles for *zmet2* or *zmet5* also result in partial reductions in CHG methylation in maize (Li et al. 2014a). *Zmet2* and *Zmet5* are expressed in similar patterns across a variety of tissues in B73 with slightly higher expression seen in *Zmet2* (Li et al. 2014a). Attempts to isolate plants homozygous for mutations in both *Zmet2* and *Zmet5* were unsuccessful, suggesting essential functions for CHG methylation in maize (Li et al. 2014a). Recent work suggests that the vast majority of "CHG" methylation in plant genomes is confined to CWG (where W is A or T) sites with very little methylation of the external C of CCG sites (Gouil and Baulcombe 2016).

CHH methylation:

There is evidence for two separate pathways for maintaining CHH methylation in plant genomes. The RdDM, involving *DRM1* and *DRM2*, plays an important role in methylation of CHH, particularly in genomic regions near genes (Law and Jacobsen 2010; Matzke and Mosher 2014). RdDM involves the production and perception of 24nt siRNAs through the combined action of two plant specific RNA polymerases, PolIV and PolV as well as RNA dependent RNA polymerase RDR2 and additional components (Matzke and Mosher 2014). The recruitment of RdDM activity to specific loci appears to require the presence of DNA methylation and specific chromatin modifications, suggesting that RdDM plays a critical role in maintaining CHH methylation patterns but may not actually represent true de novo methylation activities (Law et al. 2013; Greenberg et al. 2013; Johnson et al. 2014). True de novo methylation activities may require the activity of 21nt siRNAs with AGO6 and RDR6 to recruit DRM2 to specific target loci (Panda and Slotkin 2013; McCue et al. 2014). Arabidopsis also encodes a third domains rearranged methyltransferase, *DRM3* (Henderson et al. 2010). Interestingly, although the DRM3 protein is catalytically inactive due to changes in the active site it is a required co-factor for proper activity of DRM2 (Henderson et al. 2010).

In addition to DRM-dependent CHH methylation targeted by RdDM activities, there is also evidence for CHH methylation in deep heterochromatin that requires the chromomethylase *CMT2* (Zemach et al. 2013). These regions are likely inaccessible to PolIV/PolV activity and instead depend on CHH methylation activities from CMT2 (Stroud et al. 2014). In order to methylate these regions, CMT2 is recruited by histone methylation (Du et al. 2015). This "CHH" methylation appears to be largely confined to CWA (where W is A or T) sites (Gouil and Baulcombe 2016). Together, RdDM (utilizing DRM activities) and CMT2 maintain CHH methylation in the Arabidopsis genome.

Maize contains several DRM-like genes including *Zmet3* (*Dmt103* - Zm00001d048516), *Zmet6* (*Dmt106* - Zm00001d010928), and *Zmet7* (*Dmt107* - Zm00001d027329). *Zmet3* and *Zmet7* are retained duplicates most closely related to DRM1/2; and *Zmet6* is most similar to DRM3 (Li et al. 2014a). *Zmet3* and *Zmet7* are likely retained duplicates arising from a whole-genome duplication event in maize and exhibit similar expression patterns throughout development (Li et al 2014a). Two loss-of-function alleles have been recovered for *Zmet7* (Li et al. 2014a) but there are no documented loss-of-function alleles for *Zmet3* to date. Mutations in *Zmet7* do not have significant effects on CHH methylation in maize, but this could be due to redundancy with *Zmet3* (Li et al. 2014a).

The *Zmet6* gene encodes a protein predicted to be catalytically inactive, similar to DRM3 due to changes in the amino acid sequence near the active site of the methyltransferase domain. Maize also encodes orthologs for many of the components of the RdDM pathway (Haag et al. 2014). Several of these genes have been identified through forward genetics that identified genes required for paramutation at *R* or *Pl* (Alleman et al. 2006; Stonaker et al. 2009; Hollick 2017). Mutations in several of these genes have been shown to be required for maintaining CHH methylation at genomic regions with high (>20%) levels of CHH methylation (Li et al. 2014a; Li et al. 2015a). These mutants that

eliminate regions of high CHH methylation have relatively minor effects on gene expression in maize (Forestan et al. 2017; Anderson et al. 2018). Interestingly, unlike other grasses, maize does not contain *CMT2* orthologs (Zemach et al. 2013; Bewick et al. 2016). In maize, the deep heterochromatin regions are marked with high levels of CG and CHG methylation but low (~1-5%) levels of CHH methylation (Li et al. 2014a) that is largely confined to CWA sites (Gouil and Baulcombe 2016). It appears that this CHH methylation may depend on CHH activities of *Zmet2*/*Zmet5* (Li et al. 2014a; Gouil and Baulcombe 2016).

Demethylation*:*

While plant genomes have encoded proteins that contribute to a variety of pathways to catalyze DNA methylation, they also encode enzymes capable of active demethylation (Zhang and Zhu 2012). Demethylation is essential for certain plant developmental processes, for instance tomato fruit ripening (Liu et al. 2015) and imprinting (Bauer and Fischer 2011). Passive demethylation occurs via the failure to methylate hemi- methylated molecules that are present following DNA replication. Active demethylation (Zhang and Zhu 2012) occurs through targeted glycosylase activities. Arabidopsis includes at least four related genes including *DEMETER* (*DME*), *REPRESSOR OF SILENCING* 1 (*ROS1*), *DEMETER-LIKE 2* (*DML2*) and *DML3* that are DNA glycosylases responsible for removal of methylated cytosines through a base-excision- repair mechanism (Zhang and Zhu 2012). The maize genome encodes several DNA glycosylases (DNGs) that are homologous to those in Arabidopsis, including a DME-like gene (Zm00001d016516) and *ROS1* homologs *dng101* (Zm00001d053251) and *dng103* (Zm00001d038302) but no loss-of-function alleles for these genes have been reported.

We still have a limited understanding of the mechanisms that target these demethylation activities to specific genomic regions, but there is clear evidence that the existing methylation patterns in the Arabidopsis genome reflect a balance of methylating and demethylating activities.

Sources of variation for the maize methylome and inheritance

Understanding the frequency and distribution of differentially methylated regions (DMRs) among maize genotypes could help connect DNA methylation with phenotypic variation. In addition, understanding whether changes occur stochastically, during development, or in response to the environment is important for documenting the stability of DNA methylation. We also must understand the inheritance of variation to determine whether DNA methylation has the potential to influence heritability of traits and how to account for DNA methylation in genomic selection models or GWAS.

***Mechanisms of variation*:**

Multiple mechanisms have been proposed to give rise to variation in DNA methylation, from pure epialleles with no genetic changes to obligatory and facilitated epialleles that depend on underlying genetic variation (Richards 2006). Examples of pure epialleles (Eichten et al. 2011) and of epialleles associated with genetic changes (Eichten et al. 2012) have been reported in maize. Given that >60%% of the maize genome is annotated as transposable elements (Schnable 2009; Jiao et al. 2017), and that the composition and organization of TEs can vary greatly between inbred lines (Wang et al. 2015); this genetic variation may underpin a significant portion of variation in the methylome.

The rate of spontaneous epimutations has been studied in detail in Arabidopsis using mutation accumulation lines. Such investigations have focused on DMRs rather than single methylation polymorphisms (SMPs) because regional changes in DNA methylation are likely more functionally relevant. DMRs arise at rates comparable to genetic mutations such as SNPs (Schmitz et al. 2011; Becker et al. 2011). However, the frequency of epimutations at single cytosine residues, SMPs, is many orders of magnitude more frequent (Becker et al. 2011). It is likely different regions of the epigenome and different methylation contexts vary in SMP rates (van der

Graaf et al. 2015). Transgenerational studies in Arabidopsis highlight two significant points; SMPs can occur stochastically and SMPs are reversible, in contrast to genetic mutation. Thus, some variation in the DNA methylome arises over time through random stochastic variation. Such variation does not increase linearly with time indicating that such changes, while often stable and heritable, are also reversible. However, there was less evidence for high rates of reversible changes in methylation on a regional level (DMRs) in these studies.

Sources of variation:

Multiple studies employing a variety of technologies have demonstrated natural variation for DNA methylation in maize (Makarevitch et al. 2007; Eichten et al. 2011; Eichten et al. 2013; Regulski et al. 2013; Li et al. 2014b; Li et al. 2015b). Initial efforts identified around 700 DMRs using meDIP between B73 and Mo17 (Eichten et al. 2011). A larger scan that included ~50 diverse maize inbred lines identified 1,966 common and 1,754 rare DMRs (Eichten et al. 2013). A shift from meDIP to WGBS greatly increased the number of context-specific DMRs that were identified in maize, with 5,000 to 20,000 context specific DMRs between any two genotypes (Li et al. 2015b).

When considering this extensive epigenomic variation, it is important to consider the background genetic variation. Many DMRs can be associated with local genomic variation (Eichten et al. 2011; Eichten et al. 2013). For instance, Eichten et al (2013) reported that half of the common DMRs assessed in a panel of 50 inbred lines were associated with SNPs found within or near the DMRs; and Li et al (2015b) found that the majority of DMRs were associated with local sequence variation. These studies highlight the strong relationship between genetic and epigenetic variation. Nevertheless, examples of DMRs occurring in genomic regions that are apparently identical between inbred lines (e.g., B73 and Mo17) indicate the existence of pure epialleles (Eichten et al. 2011; Li et al., 2015b). Overall, most studies have found greater than 99% of

the methylome is conserved within a species (Li et al. 2015b). Yet, this leaves ample variation at hundreds to thousands of loci, which may contribute to phenotypic variation and breeding outcomes.

Given that DNA methylation variation can potentially occur more rapidly than genomic variation and that it is reversible, regulation of the methylome may provide a means for local and rapid acclimation or adaptation to new environments. Despite this attractive hypothesis, few concrete documented examples of environmentally induced, heritable changes in DNA methylation exist (Pecinka and Scheid 2012; Crisp et al. 2016). Profiling of maize plants subjected to heat, cold and UV revealed no evidence for consistent changes in DNA methylation in response to stress (Eichten and Springer 2015). This analysis also found that stress did not appear to increase the rate of epimutation. The examples of variation that have been identified tend to be enriched in the CHH context and lack stable inheritance patterns (Secco et al. 2015). The emerging trend that the methylome is largely impervious to environmental perturbation has important implications for breeding, allowing selection for epigenetic traits for large scale agricultural application where plants can be grown under a wide variety of environments.

Another potential source of DNA methylation variation is developmental and cellular differentiation leading to cell-type or tissue-specific variation. In animals, there are well documented examples of developmental epigenetic variation (Feng et al. 2010; Heard and Martienssen 2014). Similarly, maize endosperm and embryo have a number of differences in DNA methylation (Wang et al. 2015), consistent with findings in rice and Arabidopsis (Gehring et al. 2009, Hsieh et al. 2009; Zemach et al. 2010). In the endosperm there is widespread hypomethylation of the maternal genome, particularly at TEs, associated with the activation of endosperm specific DNA demethylases (Wang et al. 2015). Another example of a cell-type specific methylome regulation

occurs in the columella. The columella in the Arabidopsis root cap has been identified as the most hypomethylated Arabidopsis cell/tissue to date (Kawakatsu et al. 2016). Similarly, developmental regulation of DNA methylation appears to play an essential role in tomato fruit ripening, where specific gene promoters become hypomethylated during the progressive stages of ripening (Zong et al. 2013). Notwithstanding these notable examples of DNA methylation in certain tissues there is very little evidence for variation in DNA methylation between most cell types and during the majority of vegetative development (Kawakatsu et al. 2016).

In contrast to abiotic stress and development, it has long been known that tissue culture induces a remarkable degree of variation in the methylome (Kaeppler and Phillips 1993). The tissue culture process represents a traumatic genomic stress to plant cells (Phillips et al. 1994; Kaeppler et al. 2000). Despite the expectation that plants regenerated from tissue culture will be clones, regenerates often display heritable phenotypic and molecular variation (Stelpflug et al. 2014). Methylome profiles of regenerated plants have identified 479 DMRs compared to siblings not subjected to tissue culture, with a bias towards hypomethylation (Stelpflug et al. 2014). Of the hypomethylated loci, 75% reproducibly occurred in plants regenerated from independent replicate cultures and a significant number also overlap with naturally occurring DMRs (Stelpflug et al. 2014). This consistency in the genomic location of DMRs suggests that some loci are susceptible to epigenetic change in response to tissue culture. Greater than 60% of hypomethylated loci were also consistently inherited in self-pollinated progeny plants. By contrast, hypermethylated loci overall were less consistent, less reproducible in independent regenerate cultures, and less heritable. Very similar findings regarding a role for tissue culture in generating DNA methylation have been reported in rice (Stroud et al. 2013).

Inheritance of DNA methylation variation:

Genetic variation is highly heritable and exhibits well-known inheritance patterns; however, DNA methylation could be metastable (Regulski et al. 2013). The methylation state of a locus can be influenced by both *cis* and *trans*-factors (Li et al. 2014b). The combination of these factors raises the possibility of intriguing and unexpected segregation patterns of epialleles. For example, in the case of paramutation, communication of epigenetic state occurs between alleles (Chandler 2007); analysis of inheritance in epiRIL populations also suggests that allelic communication can occur at some, but not all, loci (Johannes et al. 2009; Reinders et al. 2009; Schmitz and Ecker 2012). Similarly, homologous regions located at distant genomic positions can communicate in *trans* as is the case in *PAI* silencing in Arabidopsis (Melquist and Bender 1999). Thus, efforts are ongoing to understand the prevalence and stability of a variety of known and potentially unexpected inheritance patterns.

In general, the methylation state of an allele is faithfully inherited in offspring, whether the parent is selfed or outcrossed. This is also subject to the stochastic changes and reversion that occur over time as noted above. However, both *cis* and *trans* factors can influence the methylation state of a locus, including the trans-chromosomal influence of one allele on another. For instance, when alleles with different methylation states are brought together in an F1 hybrid, *trans*-chromosomal methylation (TCM) - a paramutation like phenomena - can occur whereby the previously unmethylated loci can become methylated. In turn, this newly methylated state can be inherited in offspring, irrespective of the presence of the original methylated allele, leading to paramutation-like inheritance pattern in F2 plants (Regulski et al. 2013). This is particularly relevant in outcrossing species, such as maize, where there is also significant natural variation in DNA methylation.

Several studies have found that the majority of DMRs are stably inherited in RIL or NIL populations (Eichten et al. 2011; Regulski et al. 2013; Eichten et al. 2013; Li et al. 2014a). In many of these studies, the majority of DMRs

investigated were highly stable and exhibited locally inherited (cis) patterns, unaffected by the other allele or other genomic regions. Li et al (2014b) profiled nearly 1000 DMRs in a large set of NILs and found almost no examples of unstable inheritance. Only a small number of examples of trans-inheritance were identified, and this investigation did not identify any paramutable loci that displayed consistent characteristics of paramutation across NIL and RIL lines and qPCR validation. In part, experimental design may hamper the identification of trans-acting loci, due to the sequence similarity of interacting loci and requirement for unique alignments during sequencing read mapping in order to profile DNA methylation. Nevertheless, these investigations support the conclusion that the majority of DNA methylation variation in maize is heritable, stable and mostly controlled in cis.

Roles of DNA methylation in epigenetic phenomena and gene regulation

A primary reason for the interest in DNA methylation is its potential role as a molecular mechanism of epigenetic inheritance. Maize has historically been a model system for the discovery and genetic characterization of epigenetic phenomena including transposable element inactivation, paramutation and imprinting (Coe 2001). In addition, recent profiles of DNA methylation for multiple inbred lines of maize have revealed substantial natural variation for DNA methylation patterns that might be linked to variation in gene expression. In this section we will review the evidence for functional roles of DNA methylation in regulating gene expression in epigenetic phenomena and natural variation. Ideally, evidence for functional roles of DNA methylation might be provided through the use of mutant backgrounds or inhibitor treatments that completely abolish DNA methylation. However, there is evidence that severe reductions in DNA methylation in maize are inviable (Li et al. 2014a). Therefore, much of the available evidence for function studies is based on correlative evidence of associations or from studies of plants with minor reductions in methylation at specific contexts (Li et al. 2014a).

Transposable element inactivation:

Transposable elements (TEs) were first discovered in maize. Very early studies of TEs by McClintock and others documented variation in the activity of these elements, sometimes termed 'cycling' or transposable element inactivation (McClintock 1956; McClintock 1964). TEs with identical sequence could exist in active or inactive states. Research on maize lines derived from tissue culture found evidence for activation of several DNA transposons coinciding with reduced levels of DNA methylation (reviewed by Kaeppler et al. 2000). These studies provided strong evidence for an association between DNA methylation and transposon activity but did not show that DNA methylation was a required component for silencing TEs. Expression analyses of plants with reductions in CHH (Jia et al. 2009) or CHG methylation (Makarevitch et al. 2007) found evidence for increased transcription of a subset of transposons in the maize genome but neither study assessed the potential for functional transposon movement.

Perhaps the strongest evidence for a functional role of DNA methylation in controlling TE activity is based upon studies of TE activation in maize lines with defective RdDM machinery (Lisch et al. 2002). DNA methylation levels of Mu transposons are reduced in *mop1* plants (Lisch et al. 2002) with defective RDR2 gene (Alleman et al. 2006). Following multiple generations of self-pollination in a *mop1* genetic background there is evidence for stochastic reactivation of Mu elements (Lisch et al. 2002; Woodhouse et al. 2006a; Woodhouse et al. 2006b). These findings may suggest that RdDM activity and CHH methylation is not necessarily required for silencing of *Mu* elements, but is required for stable maintenance of the silenced state (Woodhouse et al. 2006a). Smith et al (2012) found that treatments of maize tissue cultures with the DNA methylation inhibitor 5- azacytidine could result in reactivation of another transposable element, *TCUP*. This element appears to be regulated by DNA methylation and is often reactivated during tissue culture

(Smith et al. 2012). Studies in Arabidopsis have also provided strong evidence for critical roles of DNA methylation in TE silencing using mutants that affect DNA methylation (reviewed by Underwood et al. 2017). It is likely that DNA methylation is required for the maintained silencing of TEs in the maize genome, and the low viability in genotypes with severe reductions in DNA methylation could be a direct consequence of increased TE expression and transposition.

***Paramutation*:**

Paramutation, the directed interaction between two alleles that results in a heritable change in the expression of a paramutable allele following exposure to a paramutagenic allele in a heterozygote, was first discovered at the *r1* (Brink 1956) and *b1* (Coe 1959) loci in maize. Subsequent studies have documented paramutation, or paramutation-like phenomena, at other loci in maize and other species (reviewed by Stam 2009; Hollick 2017). While the genetic sequence of the paramutated locus is the same at the paramutable locus there is a heritable change in gene expression, providing evidence for epigenetic information. At some paramutated loci there is evidence for differences in DNA methylation (Eggleston et al. 1995; Walker 1998; Sidorenko and Peterson 2001) or other chromatin marks (Haring et al. 2010). However, the exact nature of molecular mechanisms involved in establishing and maintaining paramutated states remain unclear. Genetics screens have uncovered a number of factors required for paramutation (reviewed by Hollick 2017), including components of the RdDM pathway as well as other chromatin genes, providing evidence that RdDM and/or DNA methylation is necessary for maintenance of the paramutated epigenetic state at some loci (Alleman et al. 2006; Hale et al. 2007; Barber et al. 2012). The fact that multiple components of the RdDM pathway have been isolated through forward genetic screens to find factors involved in paramutation certainly suggests a functional linkage between DNA methylation and paramutation. However, it is worth noting that only components of the RdDM pathway, not

pathways involved in maintenance of CG or CHG methylation, have been recovered. This could indicate a special role for CHH methylation or could suggest that the siRNAs produced and utilized by RdDM are critical for paramutation. Alternatively, this could be due to the fact that severe reductions in CG or CHG methylation may be inviable.

***Imprinting*:**

Imprinting (reviewed by Gehring 2013) was first characterized in maize based on differential expression of the transcription factor from the *R* locus depending upon whether this locus was inherited from the maternal or paternal parent (Kermicle 1970). Similar patterns upon parent-of-origin dependent seed color can also be observed for some alleles of the *B* locus (Selinger et al., 2001). Recent genome-wide surveys of imprinting in the maize endosperm have revealed several hundred imprinted genes in maize (Zhang et al. 2011; Waters et al. 2011). Differential methylation of the maternal and paternal alleles has been documented for several of the well-characterized imprinted genes (Haun et al. 2007; Hermon et al. 2007). Lauria et al (2004) documented evidence for extensive hypomethylation of the maternal genome in maize endosperm tissue. Based on studies in Arabidopsis and rice, where a similar phenomenon is found (Jullien et al. 2012), it is likely that this is due to expression of the DNA demethylase enzyme DME in the central cell prior to fertilization (Park et al. 2016). This global reduction of DNA methylation is then maintained following fertilization and results in reduced methylation of the maternal alleles at some loci in endosperm tissue in Arabidopsis and rice (reviewed by Gehring 2013). A genome wide analysis of DNA methylation in the maize endosperm reveals thousands of parent-of-origin DMRs (pDMRs) with many of these located near genes with imprinted expression patterns (Zhang et al. 2014). There is also evidence for histone modification differences, particularly H3K27me3, between the maternal and paternal alleles at numerous imprinted loci that may be more important for imprinting than DNA methylation (Haun and Springer 2008; Zhang et al. 2014). Interestingly,

reduced DNA methylation of the maternal allele can be associated with both maternally expressed genes (MEGs) and paternally expressed genes (PEGs) suggesting that the DNA methylation is not necessarily required for silencing during imprinting. Indeed, PEGs are more often associated with DNA methylation than MEGs (Gehring et al. 2011). In these cases, the hypomethylated maternal allele often is associated with high levels of H3K27me3 and reduced methylation may be required to allow for this other silencing mark to be added (Wolff et al. 2011; Makarevitch et al. 2013). There are also many imprinted genes that do not contain evidence for altered methylation of the maternal and paternal alleles (Waters et al. 2011; Zhang et al. 2014), suggesting that not all examples of imprinting require allelic DNA methylation differences.

DNA methylation and natural variation for gene expression:

DNA methylation could also play a critical role in generating epialleles, differences in gene expression without changes in DNA sequence. DNA methylation profiling has revealed abundant examples of natural variation for DNA methylation (DMRs) (Eichten et al. 2011; Regulski et al. 2013; Eichten et al. 2013; Li et al. 2015b). In several cases RNAseq and DNA methylation data has been collected in matched tissue samples providing an opportunity to assess potential associations between DNA methylation and gene expression levels. Eichten et al (2013) assessed the connection between DNA methylation and gene expression for 1,966 DMRs present in multiple inbred lines and located within 10kb of a maize gene and 277 examples of a significant association were documented (Eichten et al. 2013). The majority of cases reflect a negative association in which increase DNA methylation is associated with reduced or absent gene expression. Whole-genome bisulfite sequencing of 5 maize inbred lines identified a large number of context-specific DMRs in maize (Li et al. 2015b). RNAseq data on the same tissues was used to identify differentially expressed genes. A comparison of DNA methylation levels in the region surrounding the transcription start site revealed that genes with

moderate changes in gene expression (5-fold change or less) are not enriched for DMRs. However, genes with 10-fold or greater changes in gene expression are enriched for DMRs in the promoter region. Approximately 20% of genes that exhibit qualitative (on-off) differences in expression exhibit altered methylation in regions surrounding the transcription start site (Li et al. 2015b). In combination, these two studies provide evidence that DNA methylation changes are associated with some examples of natural variation for gene expression in maize and are more often found at genes with qualitative variation in expression. Makarevitch et al (2007) provided more direct evidence for a role of DNA methylation in natural variation for gene expression in maize. The *zmet2- m1* mutation, which results in reduced CHG methylation (Papa 2001; Li et al. 2014a), was introgressed into multiple genetic backgrounds and these stocks were used for expression profiling. Interestingly, the genes that are up-regulated in the *zmet2-m1* mutant lines relative to wild-type controls were significantly different in B73, Mo17 and W22. Many of these genes are expressed in wild type of some lines but silent in the others and loss of CHG methylation in the mutant results in activation of these genes. There is also evidence that natural variation for DNA methylation may result in variation in splicing patterns among different inbred lines (Regulski et al. 2013; Mei et al. 2017).

Concluding remarks

The epigenome has the potential to provide additional heritable information that can influence traits in maize and other plant species. Our ability to document the genome- wide distribution of DNA methylation in maize has provided clues to the potential for this information to influence gene expression and plant traits. Such analysis has also revealed important distinctions between Arabidopsis and maize. Continued research will be necessary to better understand the molecular mechanisms that control DNA methylation in maize and to elucidate the sources of variation for DNA methylation. It will be important to document whether substantial levels of

variation in DNA methylation are uncoupled from nearby SNPs because these will not be captured in SNP- based selection schemes. We anticipate exciting advances in our understanding of the functional relevance of DNA methylation and other chromatin modifications in maize in the coming years.

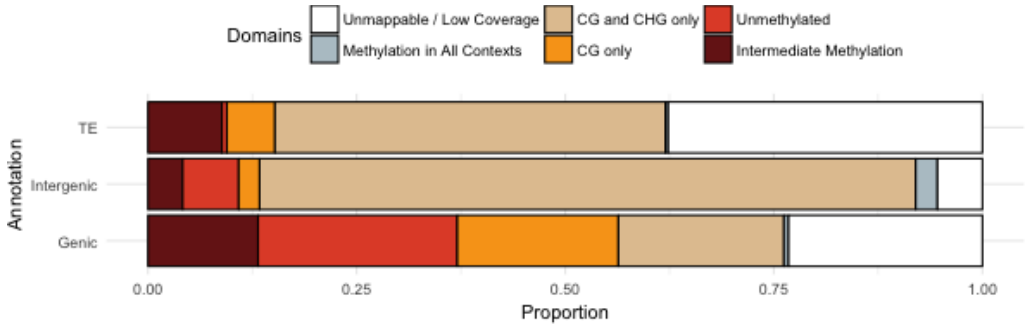

Figure 1: Frequency of methylation domains in different genomic regions. A WGBS dataset for maize earshot (Li et al. 2015abc) was mapped to version 4 of the maize B73 genome (Jiao et al. 2017). The level of DNA methylation in each sequence context was determined for each 100bp region as described in West et al., 2014. Each 100bp region was classified as genic (7.3% of genome), TE (72.3%) or intergenic (20.4%) based on B73v4 annotations. Each 100bp region was classified into one of six groups using the following criteria: Unmappable / low coverage (regions with <2X coverage), All contexts methylated (>15% CHH methylation), CG/CHG only (?40% CG and >40% CHG), CG only (>40% CG but <40% CHG), Unmethylated (<10% methylation in all sequence contexts) and intermediate methylation (sufficient coverage but not classified as one of the other groups). The proportion of 100bp regions for each subset of annotated features were determined.

CHAPTER II: Context Statement

Epigenetic variation has been observed in many plant populations. This variation can influence qualitative and quantitative traits. A key question is whether there is novel information in the epigenome that is not captured by SNP-based genetic markers. The answer likely varies depending on the sources and stability of epigenetic variation as well as the type of population being studied. We consider the epigenetic variation in several plant systems and how this relates to potential for hidden information that could increase our understanding of phenotypic variation.

Chapter II entitled ‘Literature review of an epigenetic application’ has been adapted from my work in the publication:

***Jaclyn Noshay**, Nathan Springer (2020).* Stories that can’t be told by SNPs; DNA methylation variation in plant populations. *Current Opinions in Plant Biology.*

During the course of this work several people have made contributions. Jaclyn M Noshay and Nathan M Springer collaborated throughout the brainstorming and writing process. Figures were contributed by Jaclyn M Noshay. Peter A Crisp and Robert J Schmitz provided discussion and editing suggestions. I have removed author contact information and acknowledgments as well as adapted and reformatted figures and references to be consistent throughout my thesis.

CHAPTER II:

Literature review of an epigenetic application

Stories that can't be told by SNPs; DNA methylation variation in plant populations

The availability of low-cost genetic marker technologies enabled connections between genotype and phenotype in many plant populations. These resources allow for the detection and characterization of quantitative trait loci (QTL) as well as the development of genomic prediction approaches. However, in many cases the sum of QTL effects remains well below the heritability of the trait. This so-called "missing heritability" can arise due to a variety of reasons including rare alleles, the inability to detect minor effect QTL and epistasis (Maher, 2018). Epigenetic variation, heritable variation that is not solely explained by sequence differences, could also contribute to missing heritability. We will discuss the prevalence and potential impacts of epigenetic variation in plant populations.

The term epigenetics is used in different contexts to describe both biochemical and genetic phenomena. In this review we will focus on using the term epigenetic to refer to heritable variation that is not fully linked to genetic (sequence) differences. At the molecular level epigenetic variation can be associated with differences in chromatin modification (DNA methylation, chromatin accessibility, histone variants, histone modifications), small RNAs or even protein structure (prions). DNA methylation has received the most attention as a molecular marker for potential epigenetic variation due to its relatively high heritability and the high-throughput methods for documenting genome- wide patterns (Plongthongkum et al., 2014). As we consider the potential role of epigenetics we will largely focus on studies that have evaluated variation, heritability and impacts of DNA methylation but we expect that additional studies will highlight potential roles for other molecular mechanisms of epigenetic variation.

Our desire to understand the role of epigenetic variation in plant populations stems from the potential for its contribution to phenotypic variation. Identifying the causative basis for QTL, genetic or epigenetic, increases our understanding of how variation arises and how to create or alter alleles for crop improvement. Current SNP-based QTL or GWAS approaches may fail to identify key contributors to variation of a trait if the causative chromatin modification is not captured by genetic markers. The key question we explore here is the degree to which variation in a chromatin modification such as DNA methylation will be tagged by SNPs or other genetic markers. To address this question we must consider the sources and stability of chromatin variation and the structure of the population being considered.

Sources and stability of 'epigenetic' variation:

As we seek to understand the importance of epigenetic variation in plant populations, it is necessary to consider the sources and stability of this variation. The inheritance of an epigenetic state could potentially range from high levels of stability, like a genetic variant, to complete instability from one generation to the next. Many studies that have sought to understand the sources and stability of epigenetic variation use DNA methylation in plant populations as a proxy for epigenetics. Genome-wide analyses of DNA methylation identify many differentially methylated regions (DMRs) among individuals (Schmitz et al., 2013; Kawakatsu et al., 2016; Eichten et al., 2013). While many quantitative differences in DNA methylation level for a locus have been identified, the regions changing between highly methylated and unmethylated states are most likely to represent heritable differences that may contribute to altered gene expression. It is tempting to consider all chromatin variation as true epigenetic variation, independent of sequence. However, there is abundant evidence that a significant portion of this variation in DNA methylation might be explained by genetic changes (Taudt et al., 2016). Richards (2006) created useful terminology for considering the interaction of genetic and epigenetic variation (Box 1). As we consider the potential for DNA methylation variation

to provide additional information beyond SNPs, it might be helpful to separate the sources of epigenetic variation into local (cis-acting) variants and other factors. Examples of local genetic variants that are strongly associated with an altered chromatin state (i.e., obligatory epialleles) will likely be quite stable and offer limited potential for novel information based solely on the chromatin state. In contrast, other sources of epigenetic variation including trans-acting genetic variants that trigger stable chromatin variants, environmental factors or spontaneous epimutation have the potential to create variation that is not predicted based on SNPs or other genetic markers.

Trans-acting genetic variation and genetic background can influence epigenetic variation and stability. There are well-characterized examples of genetic variants that create allelic interaction in trans (paramutation) or at unlinked genomic sites (Melquist et al., 1999). These may result in epigenetic variants that are initially predictable based on genetic variation such as SNPs. However, the behavior of this chromatin variation following segregation can create scenarios in which the epigenetic state is uncoupled with the original genetic trigger. For example, the altered epigenetic states following paramutation can be very stable or decay over the course of several generations (Hollick, 2017). Instances of genetic variants that trigger a stable trans-acting epigenetic change but are themselves lost due to segregation will create epigenetic variants that are not well- tagged by SNPs or other genetic markers.

In contrast to epigenetic changes that are triggered by trans-acting variants there are also spontaneous epigenetic changes. The best knowledge of spontaneous epimutation frequencies in plants has come from analyses of mutation accumulation lines in Arabidopsis that have minimal genetic variation (Graaf et al., 2015; Schmitz et al., 2011; Becker et al., 2011; Hofmeister et al., 2017). The rates of spontaneous epigenetic variation for single sites are several orders of magnitude higher than rates for SNPs (Ossowski et al., 2010; Schmid-Siegert et al., 2017) (Figure 1). Model based

approaches designed to estimate epimutation rates suggest some variation for different plant species but place the estimates in a generally similar range (Shahryary et al., 2020). Similar estimates were obtained for maize based on population genetics-based approaches (Xu et al., 2020). It is more difficult to estimate frequencies for differentially methylated regions but estimates suggest these occur at an overall frequency similar to site-specific methylation changes (Denkena et al., 2020). These different approaches to estimate epimutation rates in various plant species suggest slightly different rates but still place epigenetic variation into a unique position of being quite heritable but far less stable than genetic changes. Several studies have investigated whether the frequency of spontaneous epimutations may be influenced by different environments or conditions (Jiang et al., 2014; Stroud et al., 2013; Han et al., 2018; Nguyen et al., 2020; Ganguly et al., 2017; Zheng et al., 2017; Eichten & Springer, 2015; Lamke et al., 2017; Wibowo et al., 2016; Secco et al., 2015; Ji et al., 2019; Colicchio et al., 2018). While some treatments, such as tissue culture, have been associated with increased changes in DNA methylation, the influence of abiotic stresses on DNA methylation has varied from negligible to significant in different studies. The analysis of DNA methylation patterns in wild Arabidopsis populations suggested rates of accumulation of epimutations in natural environments that is similar to that observed in mutation accumulation lines, suggesting limited roles for environmental variation influencing epimutation rate (Hagmann et al., 2015). It is likely that the specific rate of epimutations and their distribution in different genomic regions may be influenced by environmental factors as well as genetic background.

Different populations: different potential for epigenetic variation

The stability of epigenetic variation and epimutation rates become critical factors as we consider the potential for untagged epigenetic variation in different types of plant populations and there are different factors that become important in disentangling genetic and epigenetic variation in these different

populations. The potential to capture DNA methylation variation through the use of SNP-based profiling will vary in different populations as a consequence of the dynamics of the age, stability and mechanisms that generate epialleles (Figure 2). We will begin with plant populations with minimal genetic variation and progress to populations with higher levels of genetic variation.

Clonal populations and inbred lines

The simplest populations from a genetic perspective will be clonal populations or inbred lines. Many crop species are clonally propagated which result in a population with relatively little genetic variation and this means that new variants (genetic or epigenetic) will not be tagged by SNPs (Figure 2A). In these species, characterizing spontaneous epigenetic variation will likely be very important for understanding sports or somaclonal variants (Latutrie et al., 2019). One prominent recent example was the discovery of the *Bad karma* locus in somaclonal variants of oil palms (Ong-Abdullah et al., 2015; Sarpan et al., 2020). This epigenetic variant arises at moderate (5-20% of individuals) frequency in clonally propagated oil palms and epigenetic assays have been developed to use for culling of affected individuals. Several recent studies have also provided insights into the epigenetic variability in some fruit species such as apple (Jiang et al., 2019; Daccord et al., 2017; Li et al., 2019), grape (Ocana et al., 2013; Xie et al., 2017) or poplar (Lu et al., 2020; Schonberger et al., 2016; Hofmeister et al., 2020). While these studies suggest promise for linking chromatin variants to novel phenotypes that have arisen in specific sports, it is worth noting that in general there is quite limited breeding progress using selection on inbred materials, suggesting limited potential for spontaneous epigenetic variants that allow for rapid shifts in quantitative traits. However, in some special cases there can be major epigenetic variation within inbred populations. The epiRILs were intentionally generated through crossing plants that are homozygous mutant for factors critical for DNA methylation with wild-type plants (Johannes et al., 2009; Reinders et al., 2009). The off-spring that are homozygous wild type (lacking mutant alleles for DNA methylation

pathways) segregate for genomic regions that have experienced loss of methylation. These populations give insight into the stability of epigenetic variants and show highly variable behavior for different loci. Some loci quickly regain wild-type methylation levels while others show stochastic rare recovery or stable unmethylated states (Catoni et al., 2017). These populations also show quantitative trait variation suggesting that segregation for varying chromatin states can influence many traits (Zhang et al., 2018; Zhang et al., 2013; Kooke et al., 2015; Cortilo et al., 2014; Furci et al., 2019). EpiRILs provide examples of how epigenetic variants that are not tagged by SNPs could result in quantitative trait variation.

Bi-parental populations

As we shift to consider populations with segregating genetic variation it becomes more challenging to disentangle genetic and epigenetic sources of phenotypic variation (Taudt et al., 2016). These populations often have much higher rates of chromatin variation, but this is present in haplotypes that also have genetic variants. Since genetic variants can have local effects on chromatin or trans effects at allelic positions (i.e., paramutation) or elsewhere in the genome, it becomes important to be able to resolve whether chromatin state differences are controlled by other genetic variants. One major challenge is that the lower stability for epigenetic variants relative to genetic variants changes the potential to use imputation approaches. While high-quality information on genetic variation from parents of a population can be accurately imputed to off-spring based on a smaller number of markers that define recombination, this approach should not be applied to high resolution maps of chromatin variants from parental genomes due to the lower stability of these variants.

Several types of genetically variable plant populations offer distinct potential for considering the role of epigenetic variation (Figure 2A). Bi-parental populations (including F2, recombinant inbred lines and near isogenic lines)

provide opportunities to monitor the stability of chromatin variants and the potential for trans-acting control of DNA methylation. In general, studies in these populations have revealed widespread evidence for relatively stable inheritance of DNA methylation levels based on the stable inheritance of epialleles with some examples of unstable inheritance (Eichten et al., 2013; Schmitz et al., 2013; Li et al., 2014; Regulski et al., 2013; Vaughn et al., 2007). While these bi-parental populations can be quite useful for insights into inheritance patterns of chromatin variation, they offer very limited ability to separate genetic and epigenetic variation. Since these populations often have limited genetic resolution each chromatin variant is often in linkage disequilibrium with many nearby SNPs or other genetic changes. Each QTL will potentially contain multiple genetic and epigenetic variants and isolating the causative variant can be challenging. In addition, the potential to tag epialleles using SNP variation will be influenced by the age of the epiallele, the stability of the epigenetic state and the mechanistic basis of the epiallele (Figure 2B-C).

Diverse association panels

Moving to diverse association panels with GWAS can provide increased genetic resolution. In these populations there are both increased numbers of crossovers as well as additional generations that provide additional opportunity for the accumulation of spontaneous epimutations. This increases the opportunity to identify chromatin variants that are not well tagged by genetic variants. To date there have been relatively few scans of chromatin profiles in very large diverse plant populations. The only comprehensive profiling of DNA methylation at true population scales has been performed in Arabidopsis (Schmitz et al., 2013; Kawakatsu et al., 2016; Dubin et al., 2015). While there are many single-base methylation polymorphisms that are detected in this population, the phylogeny based on methylation polymorphisms is highly similar to SNP- based phylogeny suggesting overall stable inheritance of DNA methylation patterns (Schmitz et al., 2013). Many of the differences in DNA

methylation in Arabidopsis populations appear to have a genetic basis with examples of local, cis-acting variants as well as trans-acting variants that frequently map to genomic locations of genes known to play a role in the regulation of DNA methylation (Kawakatsu et al., 2016; Dubin et al., 2015). More limited scans have been performed in brachypodium (Eichten et al., 2016), rice (Zhao et al., 2020), soybean (Shen et al., 2018) and maize (Eichten et al., 2013; Xu et al., 2020; Xu et al., 2019). The analysis of 45 soybean methylomes from wild accessions and domesticated lines reveals many changes in DNA methylation that are often associated with higher levels of nearby genetic variation (Shen et al., 2018). Similarly, the analysis of nearly 100 maize and teosinte methylomes identifies many differences in DNA methylation that include many examples that are associated with genomic regions that have undergone selection during domestication (Xu et al., 2020). These DNA methylation differences may have functional consequences that were the basis of selection or may simply reflect variants that 'hitch-hiked' with selection for nearby genetic variants. A capture-based profiling of DNA methylation at selected regions of the maize genome in combination with high-depth SNP panels on the same population reveals that only about half of the differentially methylated regions were effectively captured by SNPs (Xu et al., 2019). Importantly, using the DNA methylation variants could effectively predict variation for some gene expression or metabolite traits that were not strongly associated with SNPs providing evidence for potential value of DNA methylation profiles for predicting traits.

Conclusions

The analyses of diverse populations in several plant species highlight the potential for novel epigenetic variants that are not well captured in SNP-based scans to influence plant traits. Further studies will be necessary to document the full role of epigenetics in quantitative trait variation in plant populations. It will be important to design these studies in a fashion that can disentangle the effects of chromatin variation as opposed to hitch-hiking of a stable chromatin

variant with nearby genetic changes. Many of the current population genetics-based analyses of loci involved in domestication or adaptation will not have sufficient power to fully resolve the genetic and epigenetic variation at selected loci. However, in some cases these studies have pointed to intriguing potential for epigenetic variation at these loci. Several recent studies have reported potential technologies for targeted addition or removal of DNA methylation at specific loci (Gallego-Bartolome et al., 2018; Ji et al., 2018; Johnson et al., 2014; Papikian et al., 2019; Ghoshai et al., 2020). These approaches provide new opportunities for disentangling the role of DNA methylation and genetic variation by providing ways to trigger a methylation change with no genetic variation. A more complete understanding of the sources and stability of epigenetic variation in different plant populations will be critical as we seek to determine the importance, and potential value, of chromatin profiles for understanding phenotypic variation in plants.

Box 1. Spectrum of potential genetic influences on chromatin variation. Variation for chromatin state in different haplotypes can range from completely dependent on genetic variation to being completely independent. According to Richards (2006) an **obligatory epiallele** occurs when a genetic change (such as a transposon insertion or structural variant) triggers a chromatin change. For example, a transposon insertion may trigger high levels of DNA methylation for the transposon itself as well as the flanking sequences (Noshay et al., 2019). At the other extreme a **pure epiallele** would represent instances in which epigenetic variation arises with no genetic influence. Pure epialleles could arise through spontaneous epimutation or through variation triggered by environmental or developmental conditions. Between these two extremes there is the potential for several types of **facilitated epialleles**. One instance of facilitated epiallele would occur when the presence of a genetic variant, like a transposable element, predisposes a region to chromatin variation but is not completely penetrant. This leads to partial but not complete association of the genetic variant and the chromatin state. Another instance of a facilitated epiallele could occur with trans-acting effects. For example, a genetic variant that creates an inverted repeat (as seen at the PAI locus in Arabidopsis (Melquist et al., 2004) could lead to small RNAs that could trigger high levels of DNA methylation at the other allele or at other genomic locations with high homology to the inverted repeat sequence. Importantly, if the hypermethylation state is heritable high levels of DNA methylation could be maintained even after segregation of the triggering locus. This would result in an apparent pure epiallele that originally was attributed to a genetic variant. These examples highlight the complexities in determining the linkage between chromatin variation and genetic changes.

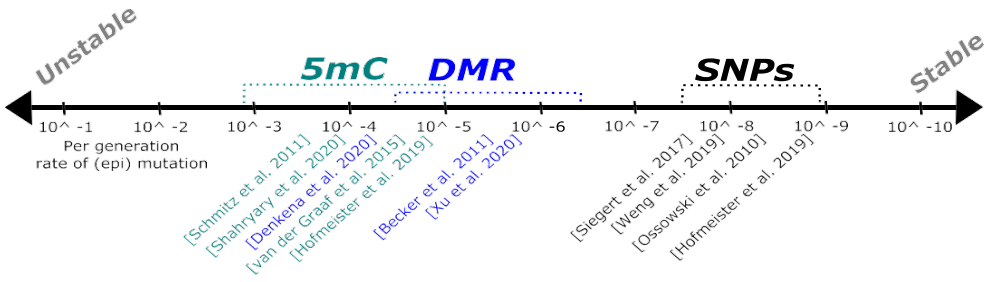

Figure 1. Relative stability of DNA methylation and SNPs. Several studies have monitored the epimutation rates for DNA methylation at specific sites (5mC - green) or for differentially methylated regions (DMRs-blue) in comparison to SNPs (black). The specific studies are referenced at the approximate position of the reported rates. Although DNA methylation levels are highly heritable they are orders of magnitude less stable than SNPs.

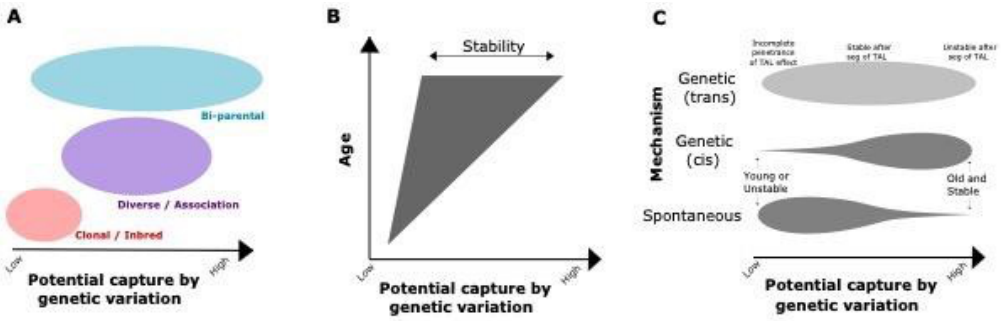

Figure 2. The potential to capture or tag epialleles using SNPs or other genetic variants is influenced by the population type, age, stability and mechanistic basis of the epiallele. A) The potential to capture epiallelic variation using genetic variation is relatively low in clonal or inbred populations. Bi-parental population and diverse association panels will exhibit a range in which some epialleles are well tagged while others are not captured. B) In all types of populations, the age and stability of inheritance for epialleles will influence the ability to capture this information using SNPs. Relatively young variants will also be difficult to tag using genetic variation scans. Epialleles that are older will exhibit a range potential capture by genetic variation resulting from different stabilities of inheritance. C) The ability to tag epialleles with genetic variation will also vary based on the mechanistic basis of the epiallele origin. Both spontaneous epialleles and cis-linked genetic causes of epialleles will have a range of potential capture depending on the age and stability. It is likely that there are greater frequencies of young, unstable spontaneous epialleles and older stable genetic (cis) epialleles. Genetic epialleles induced by a trans- acting locus (TAL) trigger will exhibit a range of potential capture by genetic variation. If the TAL trigger has incomplete penetrance for inducing a chromatin change there will be low capture by genetic variation. Loci with high penetrance of the TAL effect but unstable inheritance of the induced epigenetic state will be well tagged by genetic variation in bi-parental or association populations. However, in cases in which the TAL induces a stable effect that remains after the TAL is segregated away there will only be partial association between the TAL and the chromatin change.

CHAPTER III: Context Statement

DNA methylation and epigenetic silencing play important roles in the regulation of transposable elements (TEs) in many eukaryotic genomes. A majority of the maize genome is derived from TEs that can be classified into different orders and families based on their mechanism of transposition and sequence similarity, respectively. TEs themselves are highly methylated and it can be tempting to view them as a single uniform group. However, the analysis of DNA methylation profiles in flanking regions provides evidence for distinct groups of chromatin properties at different TE families. These differences among TE families are reproducible in different tissues and different inbred lines. TE families with varying levels of DNA methylation in flanking regions also show distinct patterns of chromatin accessibility and modifications within the TEs. The differences in the patterns of DNA methylation flanking TE families arise from a combination of non-random insertion preferences of TE families, changes in DNA methylation triggered by the insertion of the TE and subsequent selection pressure. A set of nearly 70,000 TE polymorphisms among four assembled maize genomes were used to monitor the level of DNA methylation at haplotypes with and without the TE insertions. In many cases, TE families with high levels of DNA methylation in flanking sequence are enriched for insertions into highly methylated regions. The majority of the >2,500 TE insertions into unmethylated regions result in changes in DNA methylation in haplotypes with the TE, suggesting the widespread potential for TE insertions to condition altered methylation in conserved regions of the genome. This study highlights the interplay between TEs and the methylome of a major crop species.

Chapter III entitled ‘The complex interactions of TEs and DNA methylation in maize’ has been adapted from my work in the publication:

***Jaclyn M Noshay**, Sarah N Anderson, Peng Zhou, Lexiang Ji, William Ricci, Zefu Lu, Michelle Stitzer, Peter A Crisp, Candice N Hirsch, Xiaoyu Zhang, Robert J Schmitz, Nathan M Springer. (2019).* Monitoring the interplay between transposable element families and DNA methylation in maize. *PLOS Genetics.*

During the course of this work many authors contributed. Jaclyn M Noshay, Sarah N Anderson, Peng Zhou, Lexiang Ji, William Ricci, Zefu Lu, Michelle Stitzer and Peter A Crisp performed research. In particular, Lexiang Ji, William Ricci, and Zefu Lu collected samples and performed library generation for chromatin datasets. Peter A Hermanson performed library generation for DNA methylation datasets. Sarah A Anderson and Michelle Stitzer contributed annotation and TE polymorphism tools. Jaclyn M Noshay, Peter A Crisp and Peng Zhou helped analyze data. Figures and text were contributed by Jaclyn M Noshay and Nathan M Springer. I have removed contact information and acknowledgements as well as formatted figures and references to be consistent throughout my thesis.

CHAPTER III

The complex interactions of transposable elements and DNA methylation in maize

Introduction:

Highly repetitive regions, mostly derived from transposable elements (TEs), account for the majority of DNA sequence in most crop genomes. TEs include several different types of elements (Wicker et al., 2007). Class I TEs, also known as retrotransposons, utilize an RNA intermediates and many have long terminal repeats (LTRs). Class II TEs include DNA transposons and helitrons and utilize DNA intermediates. The DNA transposons are flanked by terminal inverted repeats (TIRs). The *Zea mays* B73v4 genome contains ~300,000 structurally intact TEs (defined based on presence of structural features such as TIR/LTR and target site duplication (TSD) belonging to ~26,000 families (Jiao et al., 2017), providing opportunities to understand how TEs interact with chromatin. There are high abundance families as well as many smaller families of TEs (Jiao et al., 2017; Stitzer et al., 2019), with family sizes ranging from 1 to >15,000 in the B73v4 reference genome. These structurally intact maize TEs account for ~65% of the maize genome (Jiao et al., 2017) and estimates based on repeat masking of the maize genome, which are more sensitive for detecting fragments of TEs, suggest that >80% of the maize genome is derived from TEs (Schnable et al., 2009). Unlike the model plant species *Arabidopsis thaliana,* TEs in the maize genome are dispersed throughout chromosomes including in gene-rich regions of chromosome arms (Schnable et al., 2009; Baucom et al., 2009; The Arabidopsis Genome Initiative 2000).

TEs are a potential hazard for genome integrity as their unchecked movement will result in increasing genome size, potentially deleterious mutations and chromosome instability. Transcriptional regulatory mechanisms, including epigenetic silencing driven by DNA methylation, play important roles in the silencing of TEs (Zhang et al., 2018; Ito & Kakutani 2014; Kim & Zilberman 2014; Martienssen & Colot 2001; Yoder et al., 1997). In plant genomes, TEs are highly methylated, particularly in CG and CHG (where H is any base

except G) contexts, while genic sequences tend to have lower levels of DNA methylation – especially for CHG (West et al., 2014; Niederhuth et al., 2016; Cokus et al., 2008). As genome size increases, largely due to accumulation of LTR TEs, the overall level of CG and CHG methylation increase (Niederhuth et al., 2016). DNA methylation in plant genomes can be found in different contexts and is the result of different pathways (Noshay et al., 2018; Law & Jacobsen 2010). CG methylation is largely the result of MET1 (or orthologs) and can be propagated following DNA replication due to the presence of hemi-methylated sites. CHG methylation is largely attributable to CMT3 (or orthologs) and is targeted through a self-reinforcing loop with H3K9me2. CHH methylation occurs at asymmetric sites and is often the result of RNA-directed DNA methylation (RdDM) activities. The maize genome contains high levels of CG and CHG methylation (West et al., 2014; Regulski et al., 2013; Gent et al., 2013). CHH methylation levels are relatively low and are often found at the edges of TEs located near genes (Gent et al., 2013; Li et al., 2015).

The chromatin landscape can affect TEs insertion site preference through a variety of mechanisms (Sultana et al., 2017). However, the lack of active transposition for the majority of TEs in the maize genome has made it difficult to investigate the insertion site preference for most families. There are two well characterized active TIR families, *Mu* and *Ac/Ds*, which tend to land in regions of accessible chromatin with low methylation levels (Springer et al., 2018). In contrast, many maize LTR elements are found inserted within other TEs, which could reflect a bias towards insertion into silenced chromatin (SanMiguel et al., 1998). Family level knowledge of chromatin influences on insertion site has been limited by the lack of consistent annotations and genome-wide chromatin data sets. In addition, the current set of TE insertion sites in maize inbred lines is a result of both the insertion site preferences of TE families and the insertions that were tolerated during selection of improved varieties.

After TEs are inserted, they are likely targeted by DNA methylation and it is possible that this silencing may spread beyond the borders of the TE, resulting in changes to flanking chromatin. There is evidence that the high levels of DNA methylation targeted towards TEs can result in increased DNA methylation at flanking regions in several plant species (Hollister & Gaut 2009; Eichten et al., 2012; Quadrana et al., 2016; Choi & Purugganan 2018), potentially resulting in epialleles. These epialleles represent differences in DNA methylation levels at a genetically similar sequence in the two lines that is actually the result of a genetic change (TE insertion) nearby. These changes may represent obligatory or facilitated epialleles which require a genetic change to trigger or enable the chromatin change (Richards 2006; Springer & Schmitz 2017). There are unresolved questions about how common the spreading of DNA methylation from TEs is and whether certain families are more likely to trigger changes in nearby regions. The level of DNA methylation flanking maize LTR families is quite variable (Eichten et al., 2012). In rice, the extent of DNA methylation flanking LTR elements may be influenced by the location in the genome, age and recombination rates (Choi & Purugganan 2018). There is evidence that some *Arabidopsis* TEs can trigger changes in nearby chromatin (Hollister & Gaut 2009; Quadrana et al., 2016), but the relatively low number of elements and lack of high copy families have limited the ability to study variation at the family level that exists for post-insertional impacts using *Arabidopsis thaliana* as a model system.

The maize genome provides ample opportunities to study the interplay between TE families and DNA methylation. We assessed whether the >500 TE families with over 20 non-nested members in the B73v4 reference genome assembly exhibit variable profiles of DNA methylation in flanking regions. Three clusters of TE families were identified based on high, moderate or low levels of CG/CHG methylation in flanking regions. These patterns were found to be highly stable in other tissues and genotypes. The differences in DNA

methylation in flanking regions were associated with different profiles of chromatin accessibility and modifications within or flanking the TEs. Polymorphic TE insertion sites defined by comparison of TE content across four maize genome assemblies (Anderson et al., 2019) were utilized to monitor the likely chromatin state prior to insertion as well as the changes in chromatin that are associated with the presence of the TE. Haplotypes that lack the TE compared to those with the TE suggests that many TE families have frequent insertions within highly methylated regions, especially for LTR elements. However, a subset of TE families have a high proportion of unmethylated insertion sites. These TEs that insert within unmethylated regions frequently result in changes to DNA methylation for the flanking sequences potentially resulting in epialleles

Results:

To evaluate the interactions between TEs and chromatin we collected datasets to document genomic variation in TE content and chromatin patterns for multiple maize genotypes (B73v4, PH207, W22, and Mo17). There are 225,000 - 315,000 annotated TEs in each genome that are grouped into >23,000 families (S1 Table) (Stitzer et al., 2019; Anderson et al., 2019). DNA methylation within and near these TEs was assessed using existing or new whole genome bisulfite sequencing (WGBS) datasets (S2 Table). For each WGBS dataset the coverage and percent methylation for CG, CHG and CHH contexts was determined for 100 base-pair (bp) windows based on reads that map uniquely to the corresponding reference genome (Methods). Overall, 81% of the 2.1Gb maize B73v4 genome has at least 2X coverage but the proportion of windows annotated as TEs have slightly lower coverage (S1 Figure), likely due to the challenges of mapping to repetitive regions. The distribution of DNA methylation levels for 100bp tiles revealed bimodal distributions, especially for CG and CHG methylation, in all four genotypes (S2A Figure). Each tile was classified as unmethylated (<20%), methylated (>40%), or intermediate (20-40%) for CG and CHG methylation in each sample. The

proportion of methylated tiles varies in different genomic regions (S2B Figure). For example, the proportion of unmethylated CHG tiles varies from 1.7% in TE regions to 88.3% in exons (S2B Figure). While intergenic regions (non-TE sequences located between genes) contain some unmethylated tiles, a majority are highly methylated, similar to the profile for TEs (S2B Figure). We sought to determine whether TEs might play a role in the high level of DNA methylation observed within intergenic regions.

Different TE families exhibit distinct patterns of CG and CHG methylation in flanking regions

Previous work has classified varying patterns of DNA methylation flanking different families of LTR retrotransposons in the maize B73v2 genome (Eichten et al., 2012). This work was restricted to LTR elements and was based on a repeat-masked annotation of TEs. The availability of improved genome assemblies and annotations of intact TEs provided new opportunities to study the interplay between TE families and DNA methylation (Stitzer et al., 2019). To document the variation in the profiles of DNA methylation flanking TEs present in the B73 genome we focused our analyses on non-TE genomic regions that flank transposons and initially used DNA methylation profiles for B73 shoot tissue. Each 100bp tile was associated with the nearest TE such that regions between two nearby TEs are only assigned to the closest TE. This approach excluded the flanking regions of TEs that are inserted within other elements (nested) from our analysis (Figure 1). Given our interest in comparing the profiles of different families of TEs we focused on the subset of >500 TE families with at least 20 non-nested elements for which a robust family-wide estimate can be generated (S1 Table). This resulted in profiles of DNA methylation flanking the elements for 438 TIR families and 126 LTR families in the B73v4 reference genome. Since the orientation for many TEs, especially TIRs, is not easily determined we oriented each element based on the average level of methylation in the 5' and 3' regions such that the side with higher CG methylation, within the 1kb flanking the TE, is always aligned on the left

within metaplots. Meta-profiles of CG and CHG methylation for TIR and LTR elements exhibit different patterns, especially for regions flanking the elements (S3 Figure). Overall, there are high levels of CG and CHG methylation within the TIR and LTR elements with reduced methylation in flanking regions. The methylation decrease in regions flanking TEs is relatively gradual for LTRs while TIRs show a more distinct drop in methylation levels near the boundaries of the element (S3 Figure). CHH methylation levels are consistently low in the 1kb flanking regions for both TIRs and LTRs and therefore CHH methylation was not utilized to assess DNA methylation variation for flanking regions of TEs (S3 Figure). To assess the variability of DNA methylation patterns flanking elements in different TE families, meta-profiles of DNA methylation were generated for elements in each family with >20 non-nested members and used to perform k-means (k=3) clustering (Figure 2A-B).
Visualization of the profiles for the three clusters for LTR and TIR families revealed variable patterns for CG and CHG methylation in flanking regions (Figure 2A-B). The majority of LTR families have quite high levels of CG/CHG methylation in flanking regions with a small subset showing moderate or low flanking methylation (Figure 2A).

In contrast, TIRs have many more families with lower levels of DNA methylation in flanking regions (Figure 2B). The LTR and TIR families were classified into three categories: TE families defined by consistently high-methylation flanks (H), TE families defined by partial decay of methylation levels (which includes examples in which methylation only drops to intermediate levels or examples in which the reduced methylation does not occur until >500bp from the element) classified as moderate flanking methylation (M) and TE families defined by rapid decay of methylation for at least one of the flanking regions classified as low flanking methylation (L). Although these groups have different levels of DNA methylation for flanking sequences, they all have consistently high CG and CHG methylation levels over the TE body (Figure 2C-D).

Consistency of DNA methylation profiles surrounding TEs in multiple tissues and genotypes

These meta-profiles and classifications of TE families were entirely based upon DNA methylation data for shoot tissue of B73 seedlings. The profiles of DNA methylation are very similar in other B73 tissue types, suggesting that these patterns are stable during vegetative development (S4 Figure). We also assessed the similarity of the DNA methylation patterns for TE families in the four maize genomes. There is substantial TE presence/absence variation among these four genomes (Anderson et al., 2019) which results in different sizes and genomic distributions of TE families among genotypes. We generated heatmaps of DNA methylation profiles for 83 LTR and 318 TIR families with at least 20 non-nested members in all four genotypes (Figure 3). This revealed that TE families show profiles consistent with the B73 classification in other genomes (Figure 3), suggesting that the variability in DNA methylation profiles for different TE families is a property of the TE families themselves and not solely due to the collection of genomic locations for each family within B73.

Variable flanking methylation levels are associated with additional chromatin changes within or flanking TE families

The observation that TE families exhibit distinct patterns of CG and CHG methylation in flanking regions led us to investigate several features of the families that might be associated with this variation. Each LTR and TIR family is associated with a specific superfamily. LTRs with low flanking methylation are depleted for RLG (gypsy) families and enriched for RLC (copia) and RLX (unknown) families relative to the other groups (Figure 4A). The TIR families with moderate flanking methylation are enriched for DTC (CACTA) and DTA (hAT) families (Figure 4B). The proximity to genes for the TEs in the three groups suggests that the TEs with high levels of flanking methylation are slightly enriched in TIRs located far from genes, but there are

also many TEs within this group that are near genes (Figure 4C-D). The number of elements per family was assessed to determine if there was any enrichment for large families in high, moderate, and low flanking methylation patterns. There were no striking trends in terms of the size of TE families in the three groups, with high, moderate, and low flanking methylation groups showing a blend of family copy number (Figure 4E-F). While there are some differences in the properties of families in the groups there are no defining factors that can be used to predict the behavior of DNA methylation in the regions flanking LTR or TIR families.

The high, moderate, and low flanking methylation classifications were defined based upon patterns of CG and CHG methylation in flanking regions. We assessed how these groups varied for other chromatin modifications within and flanking these families (Figure 5). As TIR families often include many quite small non-autonomous elements and the resolution of some chromatin data can be limited, we focused on the subset of TIR elements with a length greater than 1kb (Figure 5). In contrast, the vast majority (99.97%) of LTR elements are over 1 kb and therefore we included all LTR elements. There are interesting dynamics for CHH methylation at the edges of TEs classified into the different groups. LTR families with low methylation in flanking regions show a striking peak of CHH methylation at the edges of the TE while the other classes of LTR elements have lower levels of CHH. All three groups of TIR elements exhibit an increase in CHH methylation at the edges for TE families in all three groups with the strongest enrichment in the families with low levels of methylation in flanking regions. This suggests that RNA-directed DNA methylation is most active at the edges of TEs that are located near unmethylated DNA as previously noted in maize (Li et al., 2015). The evaluation of chromatin accessibility (DNase-seq) and histone modifications (West et al., 2014; Oka et al., 2017; Zhao et al., 2018)) show differences among the TEs classified as having high, moderate or low flanking methylation. Due to the high methylation over TE bodies, we anticipated low

levels of accessibility within TE bodies. As expected, chromatin accessibility is quite low within the element itself for all types of LTR and TIR elements, with somewhat elevated levels for TIRs with low levels of flanking methylation. There are more pronounced differences in chromatin accessibility for the regions flanking the elements of families with low levels of CG/CHG methylation (Figure 5). H3K9ac and H3K56ac tend to be associated with active chromatin and would therefore be expected to have an inverse relationship with methylation trends. These histone acetylation modifications tended to be quite low within all LTR elements but showed variable levels in flanking regions (Figure 5). For TIRs, there are differences for these histone acetylation modifications within, and in flanking regions, in the three groups (Figure 5). H3K9me2 is typically associated with highly methylated silenced chromatin and is enriched within and flanking LTR elements that are classified as having high or moderate flanking CG/CHG methylation. There is less evidence for strong enrichment of H3K9me2 within TIR elements relative to flanking sequences and there seems to be a depletion of H3K9me2 in the region immediately flanking the TIRs of elements with low levels of flanking CG/CHG methylation (Figure 5). H3K27me3 is often associated with developmental silencing of gene expression and we see relatively low levels of this modification within TEs. There are higher levels of H3K27me3 in the regions flanking LTR elements that are classified as having low levels of flanking CG/CHG methylation and lower levels of enrichment for H3K27me3 in regions flanking TIR elements with moderate or low methylation for the flanking regions. Together, these observations suggest that different subsets of TE families have distinct profiles of chromatin and DNA methylation within and near the elements.

TE Expression is not strongly related to clusters defined by DNA methylation

The expression level of TE families in the different clusters was assessed to determine whether variable expression patterns could be associated with the CG/CHG methylation trends observed flanking different TE families. Many TE families exhibit detectable levels of expression in RNAseq datasets (Anderson et al., 2018). An approach to document the per-family expression of TEs that utilizes unique and multi-mapping reads (Anderson et al., 2018) was used to determine the expression level of all TE families in a panel of 23 tissues (Walley et al., 2016). We compared the proportion of TE families with detectable expression (average RPM > 1) for each of the clusters (S5A-B Figure). Among the TE families with sufficient copy number to be classified based on flanking DNA methylation levels, we see a slightly higher rate of expression for the TIR families with low flanking methylation (S5A-B Figure). However, there are also a number of TE families with high methylation for flanking regions that show expression as well (S5 Figure). For the TE families that exhibit detectable expression (115 LTR and 239 TIR families) we assessed the tissue-specific patterns of expression (S5C-D Figure). Some families in each group of high, moderate, and low flanking methylation show expression across many tissues while most families exhibit more dynamic patterns. However, there was not a clear association between clusters of elements defined by flanking DNA methylation and TE expression across tissues.

TE family level variability for DNA methylation levels at insertion sites

In the previous sections we focused on classifying TE families with >20 members inserted into low-copy regions based on flanking DNA methylation patterns. This revealed differences between TIR and LTR families and revealed clusters of TEs that exhibit differences in chromatin and TE expression patterns. This variation may be the result of differences in preference for DNA methylation level at the insertion site for TE families or due to differences in how TEs influence DNA methylation of nearby regions once they are inserted. A comparative analysis of structural annotations of TEs in the assembled genomes for four maize inbred lines including B73, PH207, W22, and Mo17

resulted in the documentation of shared and non-shared TE insertions (Anderson et al., 2019). The characterization of TE polymorphisms among these four genotypes (Anderson et al., 2019) allowed us to evaluate potential DNA methylation insertion site preferences for TE families as well as the changes in DNA methylation that accompany the presence of the TE. In this analysis we are assuming that the DNA methylation state for the haplotype lacking the TE reflects the DNA methylation state prior to insertion which is likely true in most instances. Indeed, the analysis of examples for which there is a TE insertion in one haplotype but empty sites in the other 3 haplotypes reveals that over 93% of these sites are consistently methylated or unmethylated for all three empty sites.

There are 69,292 polymorphic TE insertions among the four genotypes that have highly conserved sequence in the 200bp flanking the TE and provide the opportunity to compare DNA methylation levels in these regions without confounding flanking sequence level variation. The haplotype without the TE was defined as the "empty site" as there is no TE insertion in this haplotype but at least one other haplotype has an insertion at this site (Figure S6). The DNA methylation state for the 100bp tile containing the empty site was determined for 36,285 LTR insertions and for 16,061 TIR insertions (Figure 6A-B).

LTRs tend to exhibit a strong enrichment for high methylation at the insertion site (89.7% of empty sites are methylated) while the empty sites of TIR insertions are less often methylated (46.4% of empty sites) (Figure 6A-B). There are differences in the proportion of unmethylated empty sites for the TEs classified into high, moderate, and low flanking methylation patterns. TEs from families with low flanking methylation have a higher proportion of unmethylated empty sites while TEs from families with high flanking methylation are more frequently methylated at empty sites. This suggests that chromatin insertion site preferences may explain a large portion of the flanking methylation profiles for TE families. For all TE families with at least ten empty sites the proportion of unmethylated empty sites was assessed for each family (Figure 6C-D). All of the LTR families with high

levels of flanking methylation have >90% of the insertions located within DNA that is already methylated. In contrast, LTR families with lower levels of flanking methylation exhibit variable levels of methylation at insertion sites (Figure 6C). The TIR families exhibit more variation for the proportion of insertion sites that are methylated (Figure 6D). Since TIR DNA transposons can be mobilized through cut-and- paste transposition it is likely that a subset of the TIR empty sites may reflect excision of elements rather than the haplotype prior to insertion. It can be difficult to identify true excision sites but in many cases an excision results in elimination of the target site duplication sequence we can identify a subset of TE polymorphisms that are likely enriched for excision events. In cases in which a TE was inserted and then excised, it could influence DNA methylation of the haplotype through epigenetic memory of the chromatin marks. We assessed whether there are differences in the frequency of methylated or unmethylated empty sites for the excision sites relative to new insertions. There are not major differences in the frequency of unmethylated empty sites for the excision events compared to novel insertions (S7 Figure).

TEs can result in changes to methylation in surrounding regions

In addition to assessing methylation at insertion sites, we were also interested in documenting what happens to the chromatin at unmethylated empty sites after the TE inserts. In the haplotype with the TE insertion it is possible that the regions flanking the TE would remain unmethylated. Alternatively, the presence of the TE could be associated with an increase of DNA methylation in these flanking regions. This would result in variable methylation for conserved regions between two inbred lines that are the result of the nearby genetic change (i.e., TE insertion). The level of DNA methylation flanking the TE (the 100bp tiles on either side of the tile containing the polymorphic TE as in S6 Figure) was assessed for TEs with unmethylated empty sites (Figure 7A-B) using the TE polymorphism and DNA methylation data for all four genotypes. For the majority of the loci (54.3% of TIR insertions and 65.6% of LTR

insertions) that could be assessed, there is evidence for an increase in DNA methylation in at least one flank associated with the TE insertion into unmethylated empty sites. As expected, TEs belonging to families with low flanking methylation were less likely to be associated with gains of methylation in flanking regions relative to those with high and moderate flanking methylation. The analysis of multiple members of the same TE family revealed that 12.5% of families exhibit gains of methylation for all members of the family while there are other families with low or moderate frequencies of elements that trigger methylation gains (Figure 7C- D).

The observation that some unmethylated sites gain DNA methylation following the insertion of the TE while others do not could reflect different properties of specific insertions or could suggest a stochastic nature for methylation spreading at the edges of transposons. We looked at the patterns of methylation at 88 unmethylated empty sites in B73 with TEs present in all three of the other genomes to assess whether the patterns were consistent among genotypes. The majority (86%) of these sites gain methylation in flanking regions for at least one genotype. At most of these loci (83.3% for TIRs and 78.5% for LTRs) DNA methylation is gained in multiple genotypes, often in all three (S8 Figure). An example locus (S8 Figure) illustrates the similar gain of DNA methylation for the haplotypes containing DNA methylation. This suggests that the subset of TEs for which methylation is gained in flanking regions represent effects that are consistent across genotypes rather than reflecting stochastic variation triggered by the TE.

We investigated whether there are differences in the chromatin profiles for TEs that exhibit changes in methylation for flanking regions compared to those without changes (Figure 8). As there is only chromatin modification data available for B73 this analysis focused on the 4,791 TIR and 3,649 LTR elements that are present in B73 but have unmethylated empty sites in other haplotypes. There are not large differences in the level of CG or CHG DNA

methylation within the elements that exhibit spreading compared to those that do not (Figure 8). A visualization of DNA methylation near elements classified as spreading or non-spreading (Figure 8) suggests that changes in the average level of DNA methylation are more prevalent on one side of the element relative to the other. For elements with both flanking regions classified as unmethylated that show evidence of spreading the spreading is only observed on one flank for 54% while the remaining 46% of sites exhibit spreading for both regions. For the LTR elements we were able to assess the profiles of DNA methylation when we use the genomic orientation of the element (rather than orienting based on which side has higher DNA methylation (S9 Figure)). This reveals relatively similar differences in methylation for the 5' and 3' flanks for LTR elements. Together with the results in Figure 8 this suggests that spreading often occurs on one side of the element but that this is not defined by the orientation of the elements, at least for LTRs. The non-spreading LTR elements exhibit a stronger enrichment for CHH methylation at the edges of the elements relative to spreading LTR elements and is quite low in the regions flanking these non-spreading elements (Figure 8). The TEs without evidence for spreading of DNA methylation into flanking regions tends to have higher levels of chromatin accessibility, H3K56ac, H3K9ac and H3K27me3 in flanking regions but very little difference within the elements themselves (Figure 8). Overall, we do not see strong evidence for differences for the chromatin within the body of TEs that exhibit spreading of DNA methylation compared to those that do not but there are differences in some chromatin modifications in the regions flanking these two sets of TEs. We also investigated the attributes of TEs with, or without spreading, of DNA methylation and did not find major differences in the distance to genes, superfamily designation, family size, length, or age (S10 Figure). We also do not see any differences in the frequency of CG, CHG or CHH sites within the flanking regions or TEs bodies for the elements classified as spreading or non-spreading (S11 Figure).

Discussion

The maize genome provides opportunities to study variation in how different TE families interact with DNA methylation. In particular, the presence of abundant TEs, including many in moderate to large families, enables family level analyses of the profiles of DNA methylation and chromatin. In addition, frequent TE polymorphisms among inbred lines provides insights into the variability for both chromatin influences on insertion site preference and potential spreading of methylation following insertion. We observe distinct profiles for DNA methylation and chromatin surrounding TE families, highlighting the importance of not averaging profiles for all TEs together. It is likely that different TE families have adopted distinct strategies that enable their survival and proliferation within eukaryotic genomes which will result in distinct behaviors relative to chromatin and epigenetic regulatory mechanisms.

The TEs of the maize genome are highly methylated (Regulski et al., 2013; Yuan et al., 2002; Rabinowicz et al., 1999). This methylation likely originates from targeted *de novo* DNA methylation to transposons followed by efficient maintenance of CG and CHG methylation (Cuerda-Gil & Slotkin 2016; Bond & Baulcombe 2015; Slotkin & Martienssen 2007; Panda et al., 2016). Metaplots of DNA methylation show very high levels of CG and CHG methylation within LTR and TIR elements (Regulski et al., 2013). However, the level of DNA methylation at the edges of elements remains somewhat elevated and does not decay for several hundred base pairs. There are several factors that could influence the levels of DNA methylation near TEs. First, DNA methylation and associated chromatin modifications could influence the insertion sites for TEs. Second, TE insertions could disrupt existing chromatin and recruit DNA methylation that would spread to flanking regions. Third, the act of selection upon elements, and the chromatin changes they cause, could influence the patterns in the extant elements. We investigated each of these forces as we considered how TEs shape the methylome of maize.

TE insertion site preferences for chromatin state

The transposase or integrase enzymes of TIR and LTR transposases often include domains that can interact with histone modifications (Sultana et al., 2017). This provides the opportunity for TEs to insert within largely silenced, or active, regions of the genome. There are likely trade-offs for the TE to these strategies. Insertion within active regions likely increases the potential for TE expression and subsequent transposition. However, it can also result in higher mutation load and could allow for more efficient recognition and silencing by the host genome. In contrast, insertion within silenced regions is much less likely to result in deleterious mutations and could allow for TEs to attain very high copy number but may not allow for continued expression/mobility from these silenced sites (Sultana et al., 2017; Bennetzen & Wang 2014). Alternatively, TEs may insert at random sites into the genome and subsequent selection against insertions within genes could result in biased accumulation within silenced chromatin for extant elements.

To document the insertion site preferences for TE families it is most useful to have on- going transposition activity. This enables researchers to collect large numbers of new insertion sites and assess sequence or chromatin state enrichments represented by these loci. In maize, large numbers of insertion sites for *Mu* and *Ac/Ds* have been generated (Vollbrecht et al., 2010; McCarty et al., 2013). The analysis of chromatin accessibility and DNA methylation at these sites suggests that both of these TIR families have a strong preference for DNA that is accessible and unmethylated (Springer et al., 2018). Although LTR elements comprise the majority of the maize genome, there are no families with known on-going transposition. This has limited our ability to assess the insertion site preferences for these families. There is evidence that some plant LTR elements, such as *Tos17* in rice, can exhibit a preference for active chromatin (Piffanelli et al., 2007). However, other LTR elements seem to have a preference for inserting into other, highly methylated elements (SanMiguel et al., 1998). These preferences may even change in closely related species (Tsukahara et al., 2012).

Here we utilize site-defined TE polymorphisms between four maize inbred lines to estimate chromatin preferences for different TE families. This type of analysis is potentially confounded by two factors. First, we are assuming that the DNA methylation levels are stably inherited and that the DNA methylation state at the haplotype lacking the TE represents the ancestral state. In general, regional levels of DNA methylation are quite stably inherited (Graaf et al., 2015; Schmitz et al., 2011; Becker et al., 2011; Hofmeister et al., 2017) and it is likely that for the majority of loci, the DNA methylation at the empty site will reflect the ancestral state. In addition, we are assuming that the haplotype without the TE represents the ancestral sequence state, which is likely true for LTR elements. However, for a subset of TIR elements this could reflect a perfect excision event. Second, the extant set of TE polymorphisms reflect insertions that have occurred and not been eliminated due to selection. Insertions within highly methylated regions may be less likely to result in deleterious alleles. Therefore, the existing set of loci with polymorphic insertions may be enriched for insertions into methylated DNA. In addition, for TIR elements a subset of the polymorphisms could be the result of excision events rather than novel insertions. Despite this, we do find substantial variation among TE families. While there are many families for which all, or the majority, of empty sites are highly methylated there are also examples of families that primarily have unmethylated empty sites. In total, there are 41 families for which >75% of empty sites are unmethylated. Importantly, many of the families that are enriched for unmethylated empty sites are classified into the low-flanking methylated group. This analysis of the DNA methylation patterns at empty sites suggests that for at least a subset of TE families, particularly within LTRs, there is a preference for insertions into highly methylated DNA which limits our ability to assess the impact of insertions upon DNA methylation spreading. While the use of natural variation data to assess the presumptive chromatin of the insertion site can be difficult to unambiguously interpret, it does provide evidence for a subset of examples

where there is family specific variation for preference of insertion into methylated or unmethylated DNA.

TE influences on nearby chromatin

A subset of TE insertions will occur in unmethylated DNA. This establishes new boundaries between methylated and unmethylated DNA. Given that the majority of unmethylated DNA within the maize genome occurs within or near genes it is likely important to regulate the extent to which the TE insertion will alter chromatin of nearby sequences. Prior studies in several plant species have identified evidence that TE insertions can result in increased DNA methylation at flanking sequences (Hollister et al., 2009; Quadrana et al., 2016; Choi & Purugganan 2018; Slotkin & Martienssen 2007). This has been termed "spreading" of DNA methylation although it is not clear if this truly represents a continuous spread or simply a disruption of chromatin that allows DNA methylation to occur. In rice, the extent of spreading depends on several factors including family, age of the insertion, genomic location, and TE body methylation (Choi et al., 2018). However, we see little evidence that these variables are associated with spreading vs non-spreading elements in maize (S9 Figure). Polymorphic TEs among 140 *Arabidopsis thaliana* accessions found 50% of TEs surveyed result in DNA methylation spreading to ~300bp from the TE boundary (Quadrana et al., 2016; Stuart et al., 2016). Similarly, we find 54% of TIRs and 65% of LTRs that insert within unmethylated regions have high levels of flanking methylation.

We attempted to look at the distance of spreading but found this to be a complex issue. To assess spreading, or gain, of methylation it is necessary to focus on previously unmethylated regions. The vast majority of the maize genome is highly methylated with small patches of unmethylated DNA. Often when TEs insert into unmethylated DNA they are landing within a several hundred bp patch of unmethylated DNA that is flanked by methylation or in the region in between methylated DNA and an unmethylated gene. Insertions

into unmethylated DNA that trigger the gain of DNA methylation often result in high levels of methylation all the way to the next methylated domain. However, the extent of DNA methylation gains in regions between the insertion site and a nearby gene is much more limited. The exact mechanisms that determine whether or not DNA methylation spreads from a TE to nearby sequence are not well characterized. Previous analyses focused on maize LTRs suggest that families with high levels of CHH methylation and 24 nucleotide small RNAs have the least spreading (Li et al., 2015; Li et al., 2014. In contrast LTR families that lack CHH methylation over the LTRs are more likely to exhibit spreading. This could suggest that targeting of the RNA-directed DNA methylation machinery to TE edges could provide precise targeting and specification of boundaries. This RNA-directed CHH methylation that is found at the edge of many LTR elements may act to prevent the spread of euchromatin into TEs as well as preventing the potential spread of DNA methylation of other heterochromatin marks to flanking regions. This would be an important property that could provide a mechanism by which large, complex genomes could enable partitioning of heterochromatin and genic regions. In contrast, elements that lack RNA-directed DNA methylation may have DNA methylation maintenance and targeting mechanisms that are dependent upon histone modifications such as H3K9me2 (Jackson et al., 2002; Du et al., 2012) and these types of elements may be more likely to influence chromatin and expression of nearby genes. The proliferation of these types of elements may have more consequences for the organism. In this study, we note that there are differences for a number of histone modifications within transposable elements that have high or low levels of flanking DNA methylation.

The spreading of DNA methylation near polymorphic TE insertions can result in potential epialleles (Richards 2006; Lisch 2013; Lisch et al., 2011; Eichten et al., 2014). In these cases, flanking regions with highly similar sequence exhibit differences in DNA methylation. This difference is likely triggered by

the TE insertion but can result in differential availability of the sequence to transcription factors or machinery. Indeed, the TEs that trigger spreading account for a subset of the differentially methylated regions identified in contrasts of maize genomes (Eichten et al., 2013). Importantly, this would also predict that TE insertions that trigger spreading of DNA methylation would also have greater potential to trigger changes in the expression of nearby genes. The dynamic and potentially variable nature of the spreading could be quite important as well. In some cases of well-characterized epialleles such as *Agouti* in mice (Morgan et al., 1999), a sex-determination locus in melon (Martin et al., 2009) and *Ufo-1* in maize (Wittmeyer et al., 2018) there is evidence that a TE can lead to variable levels of spreading of chromatin that influence traits. A deeper understanding of the factors that trigger the spreading of DNA methylation and the consequences of this spread will be important as we seek to understand how TEs shape gene regulatory diversity within plant species.

Methods:

Data Availability:

Whole genome bisulfite data generated for this study is available at NCBI short read archive under accessions SRR873827 and PRJNA527657. In this study we also utilize previously published datasets that are available through the following accessions: SRR850328, SRR5436222, SRX2527280, SRR1482362 and SRR5218002.

Annotation of genes and TEs:

Whole genome assemblies for B73 (Zm00001d) (Jiao et al., 2016), W22 (Zm00004b) (Springer et al., 2018), Mo17 (Zm00014a) (Sun et al., 2018), and PH207 (Zm00008a) (Hirsch et al., 2016) were used for genome-wide analyses. All analyses were done on assemblies of chromosomes 1-10 while all scaffolds were disregarded due to the inability to assess these regions across genotypes.

Filtered structural TE annotations (Anderson et al., 2019) were used (available at https://github.com/SNAnderson/maizeTE_variation).

Polymorphic TE identification:

Identification of shared and non-shared elements was determined through pairwise comparison between four maize inbred lines (B73, W22, PH207, and Mo17). Search windows were defined by the closest, non-overlapping genes to the query TE with a syntelog in the genome being assessed. For comparison, 400bp flanking tags were extracted for each annotated TE in the genome (for each genome assessed) centered at the start and end coordinates. These flank tags were mapped to the other genomes with use of BWA-MEM (Li and Durbin 2009) in paired-end mode. Further characterization was performed on those elements with tags mapped completely within the search window.

Non-shared site-defined TEs were defined by the unique mapping of both flank tags to the window with a soft-clipped region which matches the flanking regions of the TE (does not include TE sequence). Site-defined TEs were required to maintain an absolute distance between the right and left sequence that is less than twice the TSD length of the superfamily. This resulted in a total of 69,292 non-shared site-defined elements across all pairwise comparisons used for analyses. When assessing TE polymorphisms between B73 and W22, the TSD-specific sequence was found flanking the B73 TE and the predicted W22 insertion site in 73% of cases. The analysis of identical-by-sequence genomic regions supports the high accuracy of the TE polymorphism calls (Oka et al., 2017).

WGBS:

In this study we generated novel WGBS data for B73, W22 and PH207 samples and utilized previously generated WGBS for B73 and Mo17 [SRR850328 (Li et al., 2014)] (see S2 Table for details). For B73 and PH207 shoot seedlings (slightly prior to V1) were grown for 6 days and root tissue was separated from above ground tissue (shoot) for collection of 3 biological

replicates. For WGBS DNA from the three samples was pooled and 1ug of DNA was sheared to a size of 200-300bp. These DNA fragments were then used to construct a whole-genome bisulfite sequencing library using KAPA library preparation kit (KK8232). Briefly, the DNA fragments were subjected to end repair, A- tailing, adapter ligation and dual-SPRI size selection following manufacturer's instructions. The resulting library, which has a size between 250bp and 450bp, was treated with bisulfite sodium so that unmethylated cytosines could be converted to uracil using Zymo EZ DNA methylation lightning kit (D5031). The KAPA HiFi HotStart Uracil + (KK2801) was used in the PCR reaction with the following program: 95℃/2min, 8 cycles of 98℃/30s, 60℃/30s, 72℃/4min, and a final extension step at 72℃ for 10 min. For B73 and W22 leaf tissue, plants were grown to V3 and blade tissue from the third leaf was collected for at least 2 biological replicates. DNA was pooled to generate 20ul of sheared DNA. The DNA fragments were then used to construct a whole- genome bisulfite sequencing library using the Accel-NGS Methyl-Seq DNA Library Kit (30024). Briefly, the DNA fragments were subjected to Bisulfite conversion, denaturation, adaptase, extension, and ligation following manufacturer's instructions using the Methyl-Seq Set A Indexing Kit (36024). Finally, the PCR enriched library was cleaned up using SPRI beads. The library was sequenced using Illumina HiSeq2000 with the paired-end mode and 100 cycles. The WGBS data set has been deposited into NCBI under accession numbers SRR873827 and PRJNA527657.

Trim_glore (Martin, 2011) was used to trim adapter sequences and read quality was assessed with the default parameters and paired end reads mode. Reads that passed quality control were aligned to the corresponding genome (B73v4, PH207, W22, or Mo17) using BSMAP-2.90 (Xi et al., 2009), allowing up to 5 mismatches and a quality threshold of 20 (-v 5 -q 20). Duplicate reads were detected and removed using picard- tools-1.102 ("Picard Tools – By Broad Institute" n.d.) and SAMtools (Li et al., 2009). Conversion

rate was determined using the reads mapped to the unmethylated chloroplast genome. The resulting alignment file, merged for all samples with the same tissue and genotype, was then used to determine methylation level for each cytosine using BSMAP tools. Methylation ratio for 100bp non-overlapping sliding windows across the B73v4 genome in all three sequence contexts (CG, CHG, and CHH) was calculated (#C/(#C+#T)). Each 100bp window was categorized as methylated (>=40%), intermediate (20-40%), or unmethylated (<=20%) based on the CG methylation level.

***ChIP-seq and DNase-seq data and alignments**:*

In this study we utilized previously generated chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) data including H3K27me3 (Zhao et al., 2018), H3K9ac (Oka et al., 2017) and H3K9me2 (West et al., 2014) along with novel H3K56ac data. Novel data was generated from B73 shoot tissue (described above) and ChIP-seq was performed following the general protocol of Zhang et al. (Zhang et al., 2007).

Additionally, we utilized previously generated assay for accessible chromatin with high- throughput sequencing (DNase-seq) data (Oka et al., 2017). Chromatin data was accessed from the Sequence Read Archive (SRA) database and can be retrieved through the accession numbers SRR5436222, SRX2527280, SRR1482362 and SRR5218002.

Adapter sequences were removed from raw reads using Trimmomatic version 0.33 (Bolger et al., 2014) with default setting. Qualified reads were aligned to maize B73v4 genome using bowtie 1.1.1 (Langmead, 2010) with the following parameters: -m 1 -v 2 -- best --strata --chunkmbs 1024 -S. Only uniquely mapped reads were retained, and duplicated reads were then removed using rmdup module from samtools version 0.1.19 (Li et al., 2009). Output bam files were used to count the number of reads aligning to each 100bp window of the B73v4 genome. Counts were normalized per million mapped reads (each

100bp window count was divided by one million and then by the total count across the genome).

Analysis of per-family TE chromatin modification patterns:

The analysis of per-family DNA methylation (or other chromatin modifications) was restricted to the set of TEs within families containing >=20 non-nested elements. Nested TEs are those elements with coordinates completely within another TE. There are 564 families in B73 and 401 that have at least 20 non-nested members in all four assessed genomes (S1 Table). Each 100bp window was assigned to the closest annotated TE using the bedtools closest function so that each window was only accounted for once and was only assigned to its closest TE. Although the orientation is generally known for LTR elements, it is rarely known for TIRs. In order to consistently plot trends surrounding TEs we compared the CG methylation levels in the 1kb flanks of each TE and then designate the flank with the higher methylation level to be plotted on the left (i.e., upstream of the TE). Averages were calculated by grouping TEs by their TE characterizations (order and family) and averaging within each 100bp window overlapping the TE body (normalized values) and within 1kb flanking upstream and downstream of the annotated sequence (actual distances). Relative distance was determined for the 100bp windows within the annotated TE (normalized to a 1kb window). Average CG and CHG methylation flanking TE families was used for k-means clustering for TIR and LTR plots separately using the kmeans function. The k-means clustering was performed using 2-5 clusters, but visualization of the outputs suggested the presence of three distinct clusters and the classifications were performed using a k-means = 3 clustering. Heatmaps were ordered based on clusters and the 100bp windows overlapping a TE body were collapsed and averaged. Further comparisons of the defined clusters (H, M and L) were based on analyses of average values across all members. For DNA methylation, the context specific levels of DNA methylation from WGBS for each 100bp window across the genome were utilized. For

chromatin modifications, the normalized counts per million (CPM) from ChIP-seq were calculated for each 100bp window and the average CPM across elements belonging to each category (H, M, or L) was determined. Determination of distance to genes was defined using bedtools closest with every TE being assigned to the single closest gene.

TE Expression Analysis:

Per-family TE expression was previously summarized (Anderson et al., 2018) for 23 tissues of B73 (Walley et al., 2016). Expression for each family summarized in reads per million (RPM) was downloaded from https://github.com/SNAnderson/maizeTE_variation (file: Walley_TEfamily_expression_18Jan19.txt.gz). TE families were considered expressed if the RPM value exceeded one in at least one tissue. When assessing expressed TE families across tissues, values were calculated through a log2 normalization of the family level expression for each tissue sample.

Analysis of methylation at TE absent sites and TE present flanks:

The analysis of haplotypes with and without the TE was performed based on the set of site-defined polymorphisms identified for four maize genotypes (Anderson et al., 2019). In order to have a complete list of TE insertion sites data was merged across all pairwise comparisons with every defined site in an individual genome being maintained. CG methylation levels were determined on an individual genotype basis with alignment to the corresponding genome assembly. Insertion sites were based on the 100bp window overlapping the defined site in the haplotype absent of the TE (S9 Figure). Only sites with CG methylation data in this window were considered for analyses. Sites were then classified as methylated (> 40%), intermediate (20-40%), or unmethylated (<20%) based on the genome-wide distribution of CG methylation. When assessing family-based insertion patterns, the subset of 193 TE families with at least 10 insertions with DNA methylation data were considered (S4 Table). For haplotypes present for the TE, the flanking methylation was determined based

on the 100bp windows on either side of the TE, but not overlapping the TE coordinates (S9 Figure). A single classification was made based on the average CG methylation for these flanking windows. When identifying family-based proportions of spreading, the subset of 150 TE families with at least 4 survey-able unmethylated insertion sites were considered (S4 Table).

Tables:

Table S1: TEs and TE families

			B73	W22	PH207	Mo17
Transposable Elements	TIR	Genome	172,840	154,914	131,643	140,139
		>20 non-nested member family	49,585	48,220	52,618	51,048
	LTR	Genome	142,190	136,433	93,524	137,762
		>20 non-nested member family	50,882	50,672	32,768	46,816
TE Families	TIR	Genome	1,218	1,183	925	1,187
		>20 non-nested members	438 (318 shared*)	419	400	424
	LTR	Genome	23,459	22,931	23,757	23,335
		>20 non-nested members	126 (83 shared*)	142	134	140

* Shared TEs are those that are annotated across all 4 genotypes (B73, W22, Mo17, and PH207)

Table S2: Whole Genome Bisulfite Sequencing Metadata

Sample ID	Data Location	Ref	Samples	Alignment Genome	Reads	Mapping Rate	Tiles with Data	Tiles with Coverage
B73 Shoot	SRR8738272	NA	1	B73	636,294,670	0.65	0.95	0.81
B73 Root	SRR8740850	NA	1	B73	230,772,157	0.65	0.94	0.79
B73 Leaf	SRR8740851	NA	6	B73	928,696,652	0.62	0.96	0.76
Mo17 Leaf	SRX3311637	Li et al., 2014	1	Mo17	141,635,754	0.63	0.91	0.7
PH207 Shoot	SRR8740852	NA	1	PH207	430,997,520	0.65	0.99	0.98
W22 Leaf	SRR8740853	NA	2	W22	333,172,392	0.63	0.93	0.79

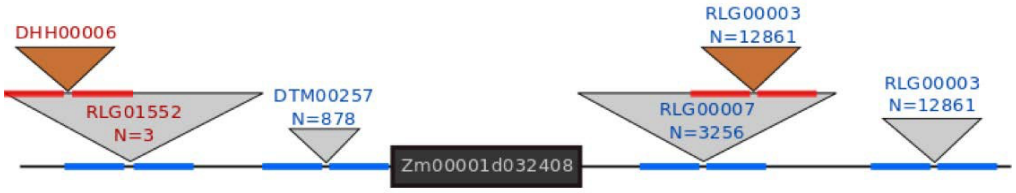

Figure 1: A region on chromosome 1 of the B73v4 genome from 225,749,884bp to 225,830,165bp is displayed as a schematic of genes (rectangles) and TEs (triangles). All TEs are labeled with their TE family name and family size with red text indicating small families (< 20 members) and blue text indicating large families (>= 20 members). Nested (within another TE) and non-nested elements are shown in orange and grey respectively. Flanking regions are identified by color based on whether they are outside of other TEs (blue) or located within other TEs (red). Flanking regions within other TEs are excluded from our analyses.

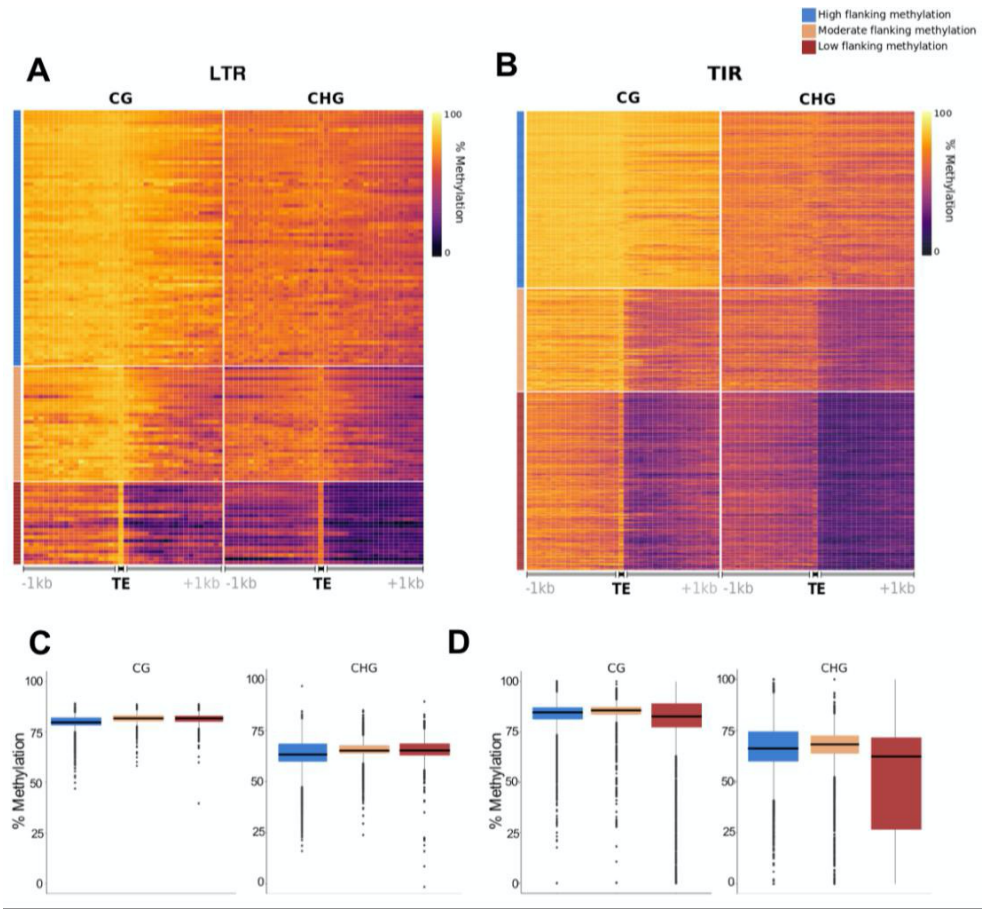

Figure 2: Variation in the profiles of CG and CHG methylation flanking TE families. (A- B) For 126 LTR families (A) and 438 TIR families (B) that include >= 20 non-nested annotated members the metaprofile of CG and CHG methylation was determined. Each element was oriented such that the 1kb flanking the TE with higher average methylation is plotted on the left. The profiles of CG and CHG methylation were used to perform k- means (n=3) clustering and the profiles were plotted with a heat map. Three clusters were defined using k-means clusters of CG and CHG methylation profiles for each element and these clusters are classified as having high (H, blue), moderate (M, orange) and low (L, red) flanking CG and CHG methylation are indicated. (C-D) The average level of CG and CHG methylation within each LTR (C) and TIR (D) assigned to the clusters was determined and used to generate boxplots to investigate differences in methylation levels for the three clusters.

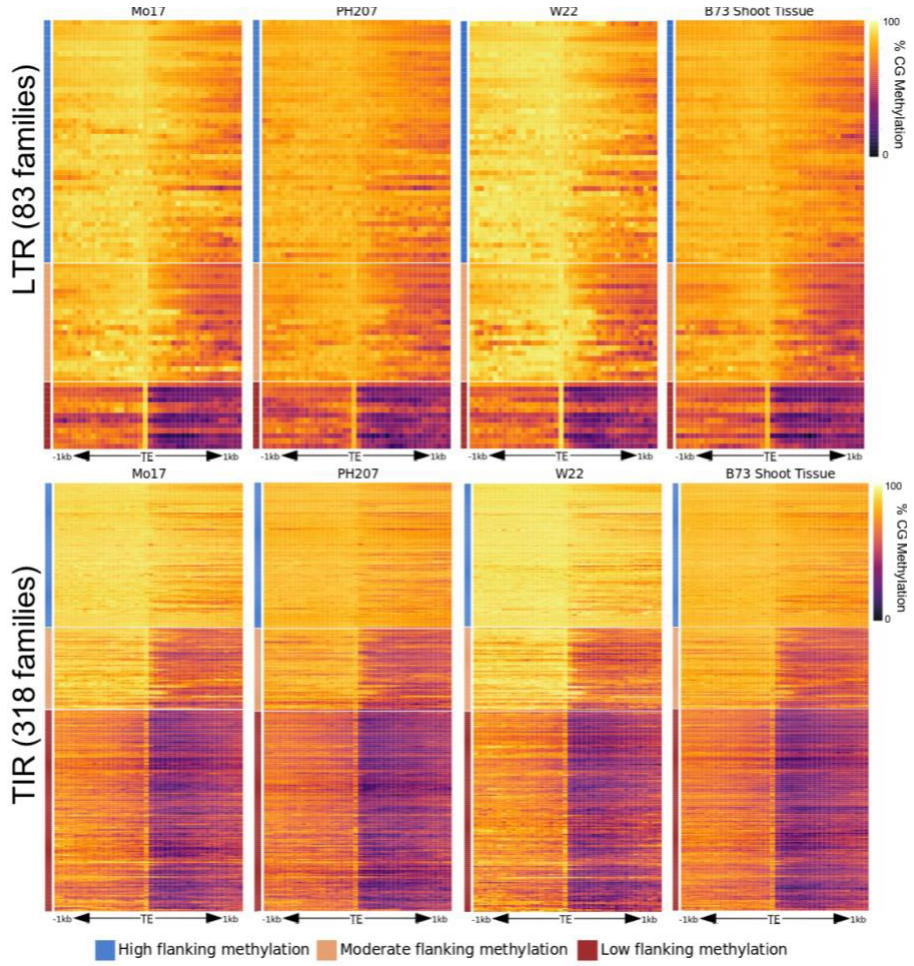

Figure 3: Consistency of methylation profiles surrounding TEs in different maize inbred lines. For 83 LTR families and 318 TIR families that have at least 20 non-nested members in all four genotypes the metaprofile of CG DNA methylation was determined. The order of families was kept the same as in Figure2, although a subset of the families were omitted as they did not have 20 members in at least one other genotype. The DNA methylation levels were determined based on the alignment of WGBS to the genome assembly from which it was derived and using the elements annotated within that genome. Very similar consistent patterns are also observed using CHG methylation profiles.

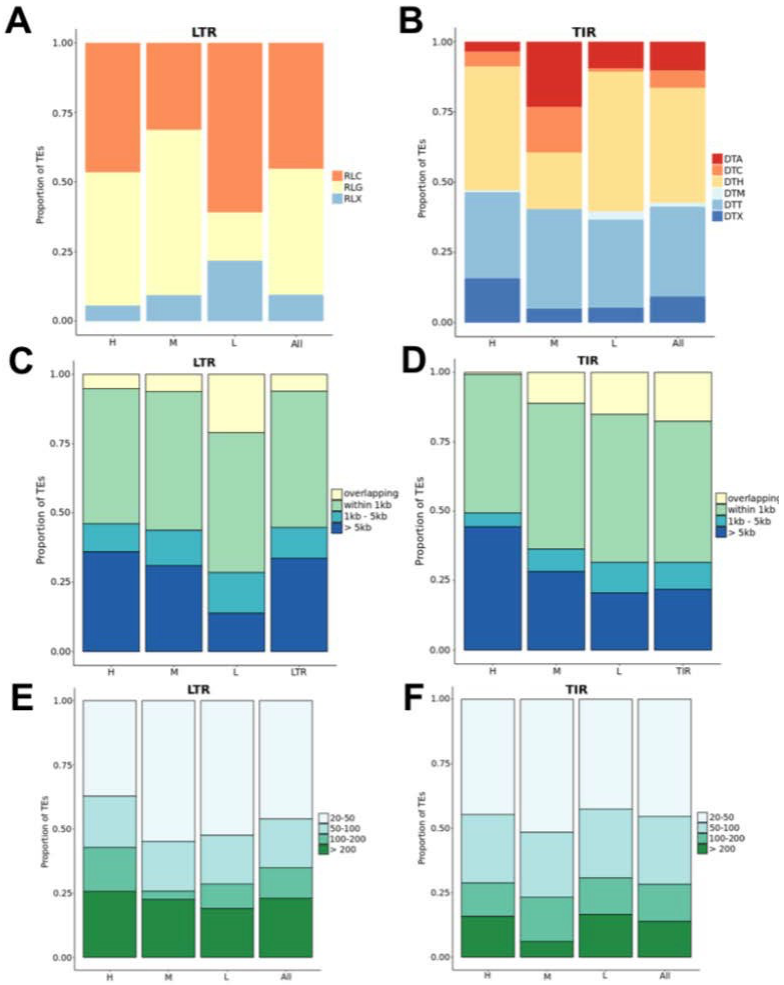

Figure 4: Analysis of attributes for members of TE families of high, moderate and low flanking CG/CHG methylation. (A-B). The proportion of superfamily designations for TEs classified as having high, moderate or low flanking methylation were compared to the proportion for all TE families. LTR elements (A) are classified into copia (RLC), gypsy (RLG) or unclassified (RLX) superfamilies. TIR elements (B) are classified as hAT (DTA), CACTA (DTC), PifHarbinger (DTH), Mutator (DTM), Tc1/Mariner (DTT) or unclassified (DTX). (C-D) For all elements of TE families classified as having high, moderate or low flanking methylation the distance to the nearest gene was determined and binned overlapping, within 1kb, 1-5kb or >5kb. The proportion of elements with varying proximity to genes was compared for the different clusters of TEs relative to all LTR or TIR elements. (E-F) A similar analysis was done to compare the proportion of families with different copy numbers in each cluster to all LTR (E) or all TIR (F) elements. Family size was classified into 4 groups: 20-50 members, 50-100 members, 100-200 members, and >200 members from light to dark color respectively.

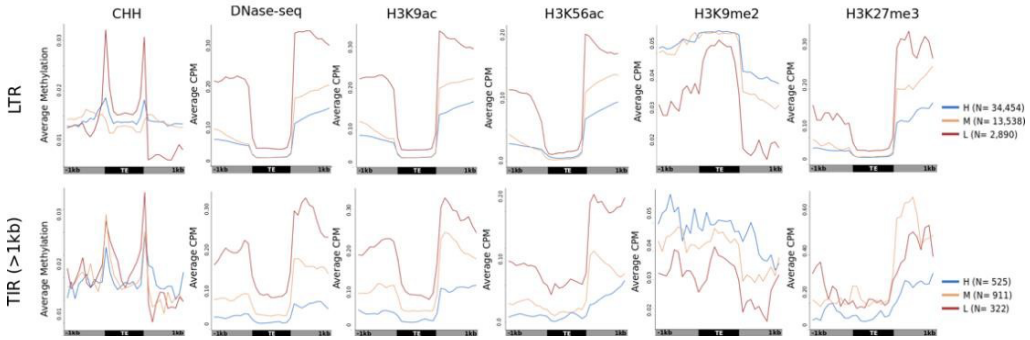

Figure 5: Analysis of chromatin within and surrounding the three clusters of TEs defined based on flanking levels of CG and CHG methylation. The relative abundance of CHH methylation, chromatin accessibility (DNase-seq and histone modifications H3K9ac, H3K56ac, H3K9me2 and H3K27me3 were determined within and flanking TEs that were classified as having high, moderate or low flanking methylation patterns. The CHH methylation was determined for 100bp tiles and the average level within the H, M and L categories was calculated. For the remaining chromatin marks (ChIP and DNase), the average CPM for each 100bp bin within the categories H, M and L was calculated. The top set of plots show the metaprofiles for LTR elements (N=50,882) in the three clusters with high (blue), moderate (orange) and low (red) indicated with different colors. The lower set of plots show the metaprofiles for the three clusters of TIR elements (N=1,758) only using elements >1kb in length. Number of elements in each group indicated next to key.

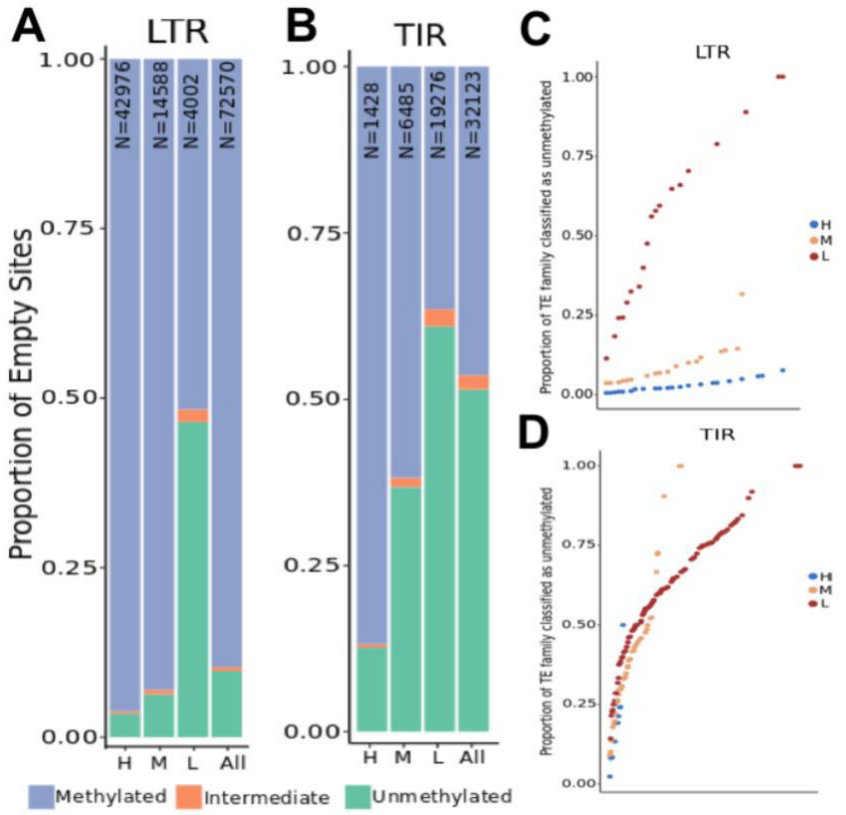

Figure 6: Levels of CG DNA methylation at empty site haplotypes. There are 75,570 LTR (A) and 32,123 TIR (B) TE polymorphisms for which there is WGBS data for the haplotype that lacks the insertion (empty site). The empty sites were classified based on CG methylation level as unmethylated (<20%), intermediate (20-40%), or methylated (>40%) and the distribution of these three groups is shown for all TEs as well as TEs in families classified as having H, M or L flanking methylation. (C-D) For each TE family with at least 10 empty sites the proportion of CG unmethylated empty sites was determined and used to rank order the families.

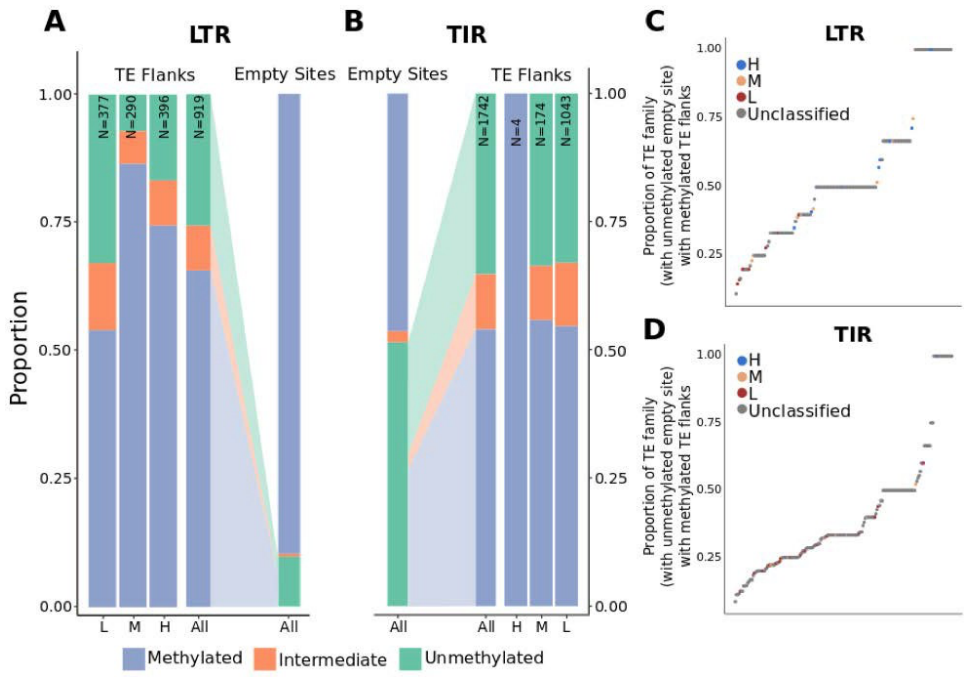

Figure 7: Analysis of CG DNA methylation changes induced by TEs. (A-B) The subset of TEs that are located within CG unmethylated empty sites could assessed for changes in levels of CG methylation in flanking regions. There are 919 LTR elements (A) and 1742 TIR elements (B) that represent insertions into unmethylated empty sites for which there is CG methylation data for the regions flanking the TE. The proportion of these TEs that show methylation, based on the average CG methylation of the TE flanks, was determined for all of these sites as well as the subsets that are near TEs belonging to families classified as having H, M or L flanking CG and CHG methylation. For the set of LTR (C) or TIR (D) families that have at least four insertions into CG unmethylated regions, the proportion of family members that gain CG DNA methylation was determined and used to rank order the families. Each family was color coded based on its classification for H, M or L flanking methylation. The unclassified families did not have enough elements to be classified as H, M, or L.

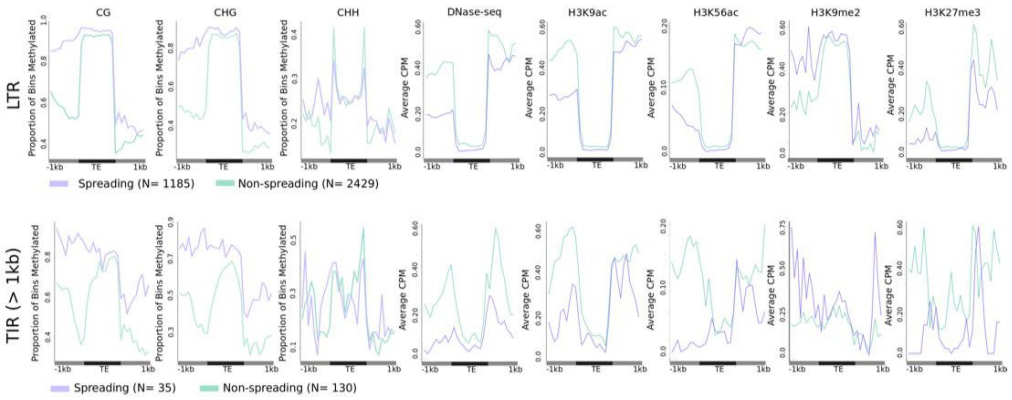

Figure 8: Chromatin profiles at elements with or without spreading of DNA methylation. For LTR or TIR elements that are inserted into CG unmethylated empty sites we assessed the chromatin profiles based on the proportion of bins methylated (>40% for CG/CHG and >2% for CHH) or average CPM (ChIP-seq and DNase-seq) for elements with spreading (purple) or without spreading (green) of DNA methylation in B73. Elements are oriented with the highest average level of methylation on the left. Number of elements in each group is indicated next to the key.

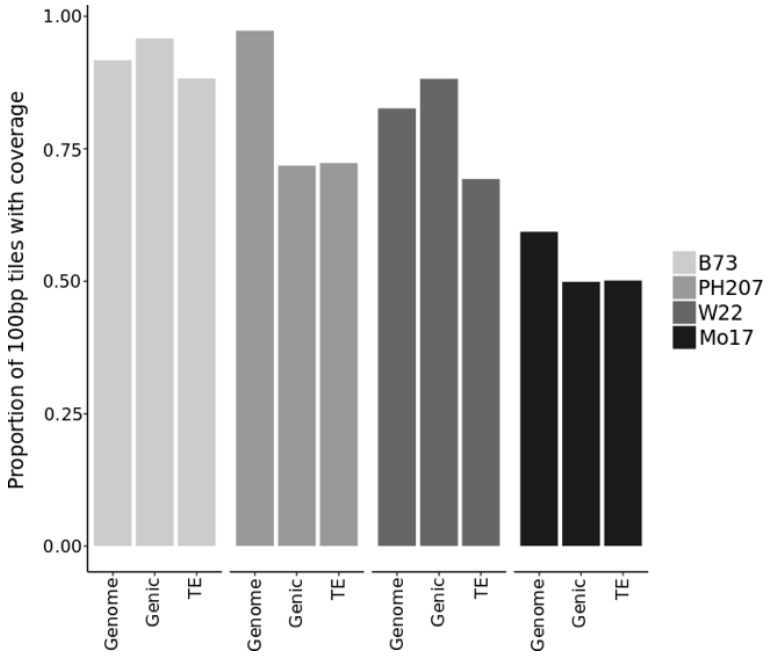

Figure S1: The proportion of 100bp tiles with >2X coverage for WGBS data in B73_Shoot, PH207_Shoot, W22_leaf and Mo17_leaf (left to right) when mapped to the corresponding genome was determined. For each dataset/genome the proportion of 100bp tiles with >2x coverage for all regions (left), only genes (middle), and only TEs (right) is shown.

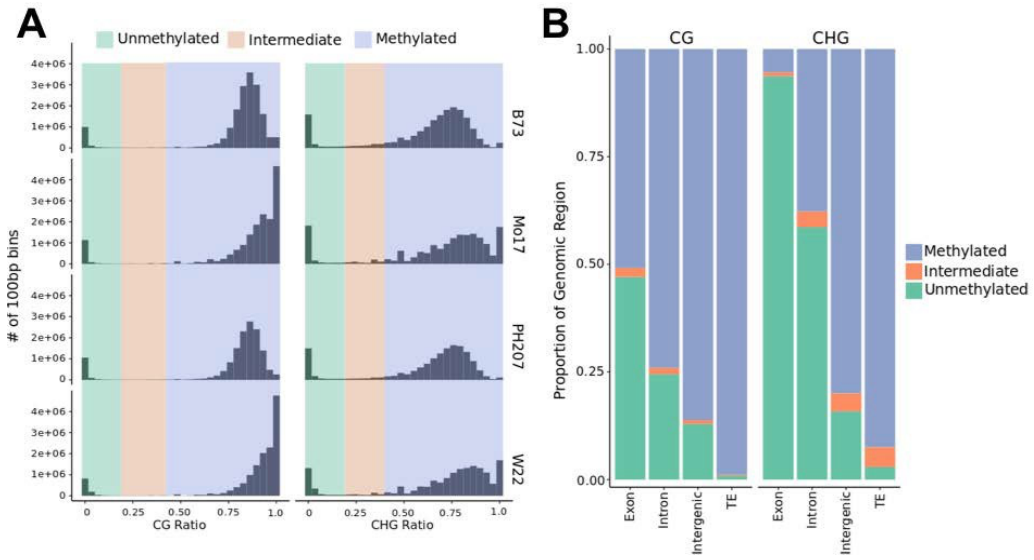

Figure S2: A WGBS dataset for maize genotypes (B73, W22, PH207, and Mo17) was mapped to the corresponding genome. (A) The level of DNA methylation in each sequence context (CG and CHG) was determined for each 100bp tile and histograms of CG and CHG DNA methylation are shown with classifications of methylated (purple), intermediate (orange), and unmethylated (green) regions indicated. B) All 100bp tiles were classified as TE, exon, intron, or intergenic based on B73v4 annotations. Each 100bp tile within these regions were classified into as methylated (>= 40%) intermediate (methylation levels > 20% and <40%) or unmethylated (methylation levels <20%). The proportion of 100bp tiles classified as methylated, intermediate or unmethylated in each type of annotation was determined.

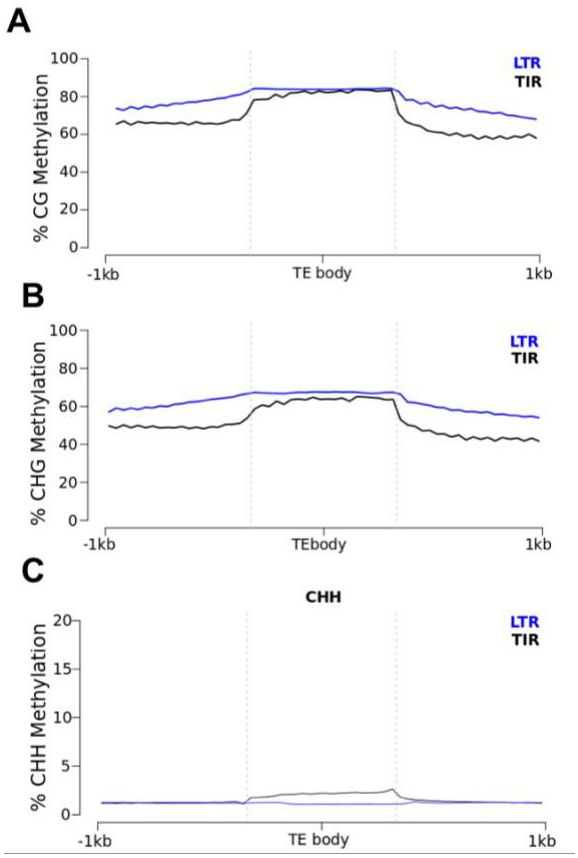

Figure S3: The DNA methylation levels within and 1kb on either side of TIR and LTR elements in the maize genome was assessed. Each 100bp bin was assigned to the closest annotated TE in the B73v4 genome. The average DNA methylation levels for LTR (blue) and TIR (black) elements is shown in the CG (A), CHG (B) and CHH (C) contexts. Internal regions of TEs were normalized to a length of 1kb and are marked by vertical dotted lines.

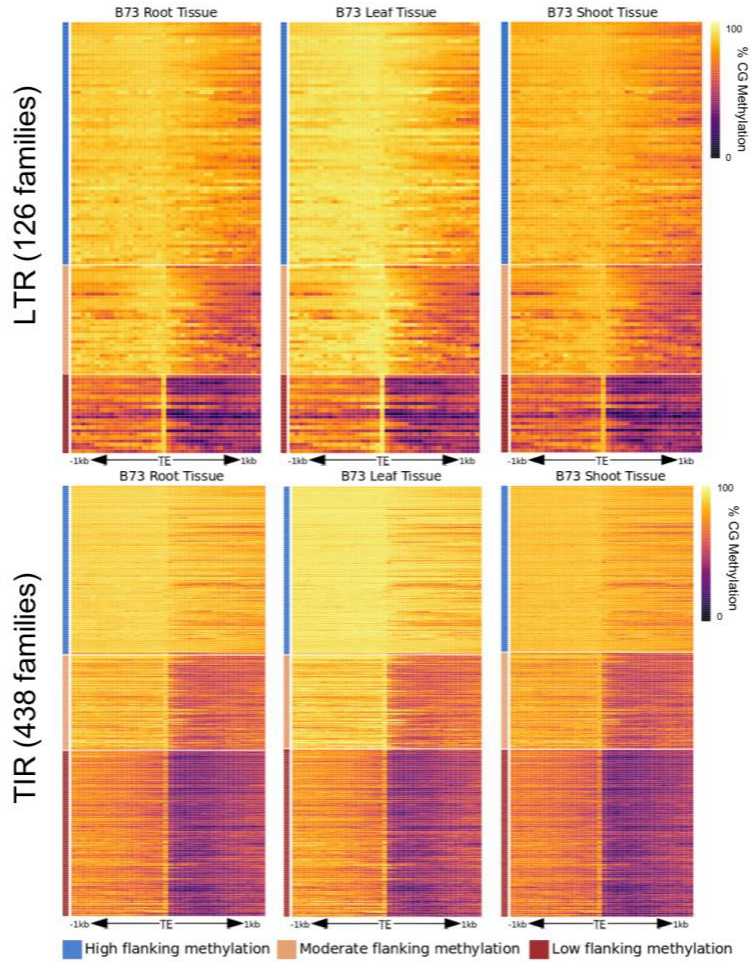

Figure S4: Consistency of methylation profiles surrounding TEs throughout vegetative development. Per-family CG methylation profiles were determined as in figure 2 using WGBS data from two additional tissues of B73 (B73 root and B73 leaf) and compared to the profiles for B73 shoot tissue. The heatmaps retain the same clustering order that was determined for B73 shoot data (as in Figure 2). Very similar consistent patterns are also observed using CHG methylation profiles.

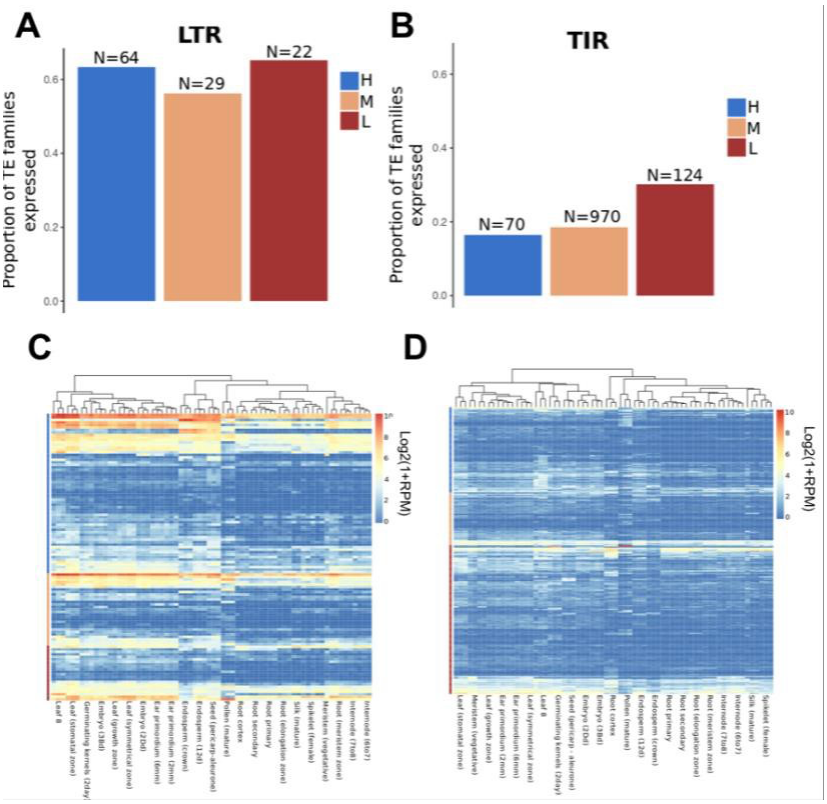

Figure S5: Expression of TE families in three clusters defined as having high (H), moderate (M) or low (L) levels of CG and CHG methylation in flanking regions. The per-family expression level for each TE family was determined in a panel of 23 tissues of B73 using RNAseq data. The proportion of LTR (A) or TIR (B) families that were classified as high (H), moderate (M) or low (L) flanking methylation levels that have detectable (average RPM > 1) expression is shown. The number indicated above each bar represents the number of families expressed within each group. (C) For LTR TE families with detectable expression (average RPM > 1) a clustering was performed based on the log2 of the family level expression for that tissue sample. (D) A similar analysis is of TE expression is shown for TIR families.

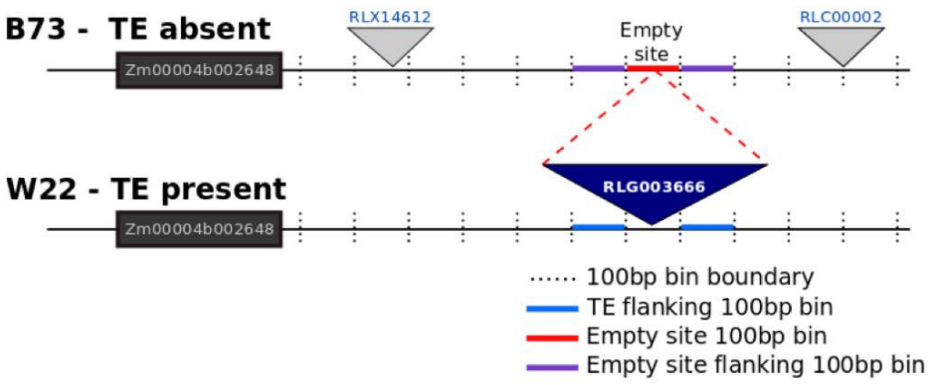

Figure S6: A region on chromosome 1 of the B73v4 and W22 genomes is displayed as a schematic of genes (rectangles) and TEs (triangles). All TEs are labeled with their TE family. The blue TE is representative of a site-defined polymorphic TE where the insertion is in the W22 genome. The dashed red lines indicate the location of the empty site in B73. 100bp bins are represented with black dashed lines along the chromosome. These bins were used for analysis of methylation and chromatin data across genomes. The red line indicates the 100bp bin representative of the empty site while the purple and blue lines indicate the flanking 100bp bins of the empty site and TE, respectively.

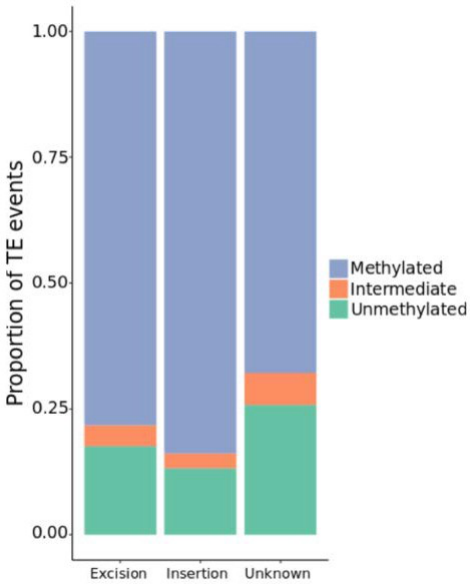

Figure S7: Identification of excision events. The target site duplication (TSD) was identified for each empty site and the corresponding TE "insertion" was assessed for presence of the left and right TSD. If the sequence of both TSDs were identical to the insertion site, the event was classified as an insertion event (N = 4707). If neither TSD sequence matched, the TE was classified as a putative excision event (N = 579), and TEs with one matching TSD sequence were classified as unknown (N = 1158). The relative proportion of these groups identified as unmethylated, intermediate, and methylated was determined.

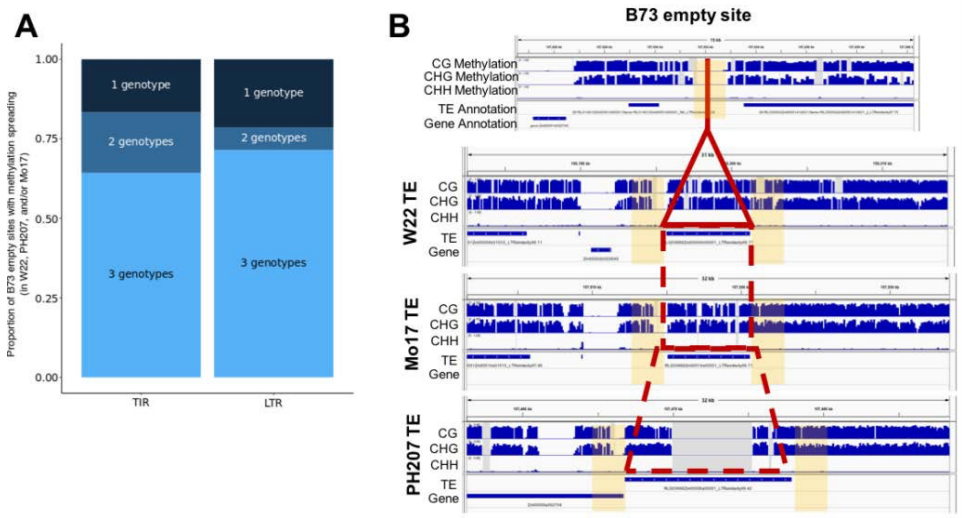

Figure S8: Consistency of DNA methylation changes near TEs. A subset of 613 TEs identified as absent in B73 (empty site) and present in W22, PH207, and Mo17 were identified. These included 88 examples in which the B73 empty site is unmethylated and 76 of these have flanking methylation gains in at least one other genotype. (A) The proportion of these 76 gains that are observed in 1, 2 or 3 genotypes was determined and plotted for TIRs and LTRs, respectively (A). (B) An example of a locus on chromosome 1 with an unmethylated empty site in B73 and gains of methylation in flanking regions for the other three genotypes is shown.

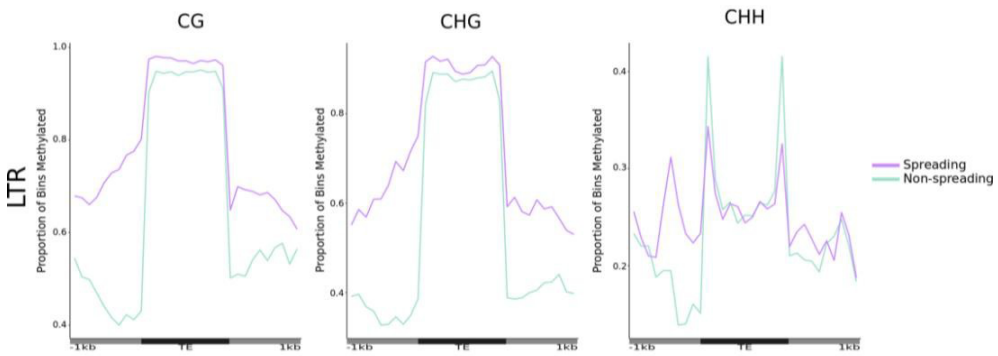

Figure S9: CG, CHG, and CHH DNA methylation profiles at LTR elements with or without spreading of DNA methylation. For these plots the orientation is based on the annotation of the element (rather than the level of DNA methylation in flanking regions). For elements that are inserted into unmethylated empty sites we assessed the DNA methylation profile based on the proportion of bins methylated (>40% for CG/CHG and >2% for CHH) for elements with spreading (purple) or without spreading (green) of DNA methylation in B73.

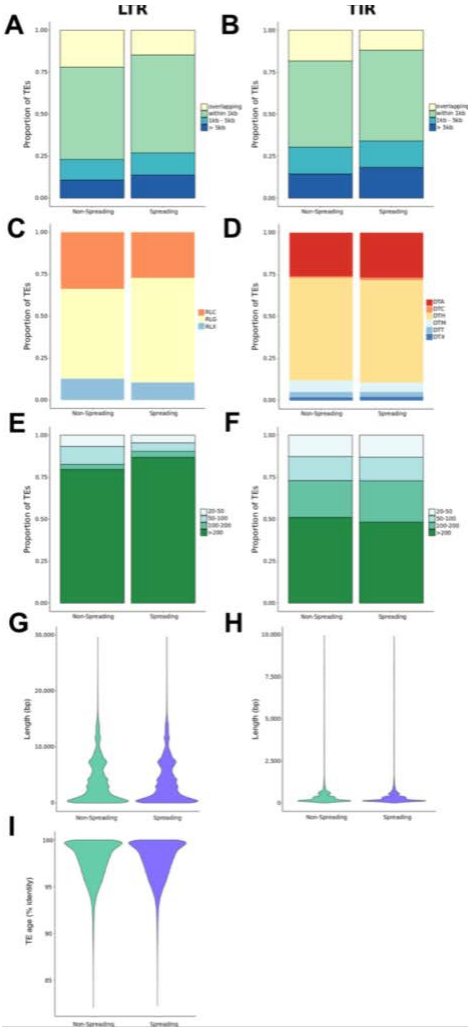

Figure S10: Analysis of attributes for members TEs classified as spreading or non- spreading. The proportion of superfamily designations for TEs classified as spreading or non-spreading. (A-B) For all elements classified as spreading or non-spreading the distance to the nearest gene was determined and binned overlapping, within 1kb, 1-5kb or >5kb. The proportion of elements with varying proximity to genes was compared for the spreading and non-spreading elements within TIRs and LTRs. LTR elements (C) are classified into copia (RLC), gypsy (RLG) or unclassified (RLX) superfamilies. TIR elements (D) are classified as hAT (DTA), CACTA (DTC), PifHarbinger (DTH), Mutator (DTM), Tc1/Mariner (DTT) or unclassified (DTX). (E-F) A similar analysis was done to compare the proportion of families with different copy numbers in spreading and non-spreading groups of LTR (E) or TIR (F) elements. Family size was classified into 4 groups: 20-50 members, 50-100 members, 100-200 members, and >200 members from light to dark color respectively. (G-H) The length of LTR and TIR elements was compared between the spreading and non-spreading groups. (I) For LTR elements the distribution of LTR similarities (% sequence identity) is shown for spreading and non- spreading elements. The greater % identity indicates younger age of the elements.

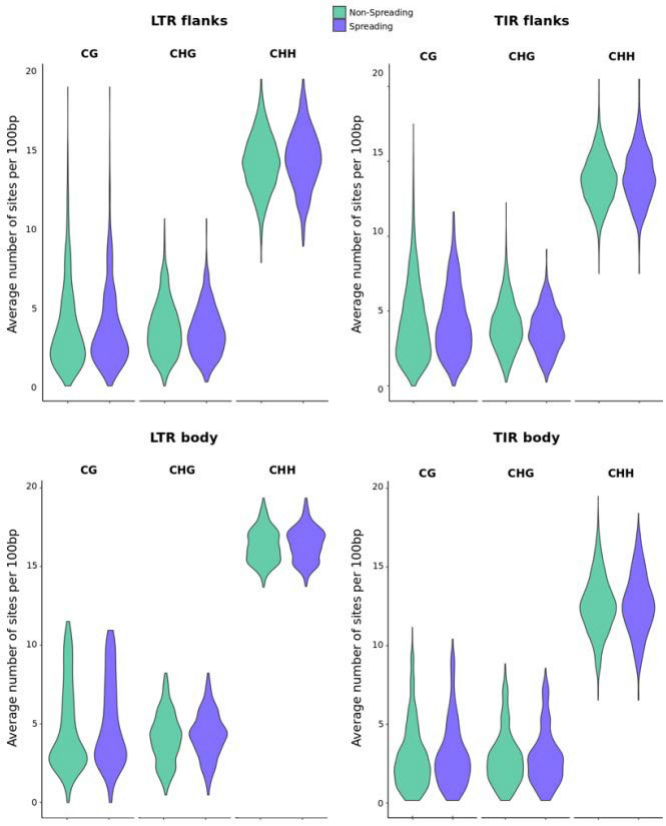

Figure S11: Analysis of CG density for spreading (purple) and non-spreading (green) for LTRs (left) and TIRs (right). The average number of cytosine sites per 100bp window for the 1kb flanking (top) and TE body (bottom) in each context (CG, CHG, and CHH) was calculated and the proportion of cytosines was plotted for each TE

CHAPTER IV: Context Statement

Transposable elements (TEs) have the potential to create regulatory variation both through disruption of existing DNA regulatory elements and through creation of novel DNA regulatory elements. In a species with a large genome, such as maize, the many TEs interspersed with genes creates opportunities for significant allelic variation due to TE presence/absence polymorphisms among individuals. We used information on putative regulatory elements in combination with knowledge about TE polymorphisms in maize to identify TE insertions that interrupt existing accessible chromatin regions (ACRs) in B73 as well as examples of polymorphic TEs that contain ACRs among four inbred lines of maize including B73, Mo17, W22, and PH207. The TE insertions in three other assembled maize genomes (Mo17, W22 or PH207) that interrupt ACRs that are present in the B73 genome can trigger changes to the chromatin suggesting the potential for both genetic and epigenetic influences of these insertions. Nearly 20% of the ACRs located over 2kb from the nearest gene are located within an annotated TE. These are regions of unmethylated DNA that show evidence for functional importance similar to ACRs that are not present within TEs. Using a large panel of maize genotypes, we tested if there is an association between the presence of TE insertions that interrupt, or carry, an ACR and the expression of nearby genes. While most TE polymorphisms are not associated with expression for nearby genes the TEs that carry ACRs exhibit an enrichment for being associated with higher expression of nearby genes, suggesting that these TEs may contribute novel regulatory elements. These analyses highlight the potential for a subset of TEs to rewire transcriptional responses in eukaryotic genomes.

Chapter IV entitled ‘Transposable element influence on maize regulatory regions’ has been adapted from my work in the publication:

***Jaclyn M Noshay**, Alexadre P Marand, Sarah N Anderson, Peng Zhou, Maria Katherine Mejia Guerra, Zefu Lu, Christine O'Connor, Peter A Crisp, Candice N Hirsch, Robert J Schmitz, Nathan M Springer. (2020).* Assessing the regulatory potential of transposable elements using chromatin accessibility profiles of maize transposons. *GENETICS.*

During the course of this work many authors contributed to data collection, processing, analysis, and writing. Jaclyn M Noshay, Alexandre P Marand, Sarah N Anderson, Peng Zhou, Maria Katherine Mejia Guerra, Zefu Lu, Christine O'Connor, and Peter A Crisp performed research. In particular, I helped process samples and perform data analysis. Zefu Lu and Alexandre P Marand provided computational tools for accessible chromatin analysis. Maria Katherine Mejia Guerra provided eQTL analyses. Sarah N Anderson and Christine O'Connor provided analyses for assessment of polymorphic TEs. Peng Zhou and Peter A Crisp provided computational support. Figures were contributed by Jaclyn M Noshay and Alexandre P Marand. Texts were contributed by Jaclyn M Noshay, Nathan M Springer with editing support from Robert J Schmitz and Candice N Hirsch. I have removed author contact information and acknowledgements as well as formatted figures and references to be consistent throughout my thesis.

CHAPTER IV

Transposable element influence on maize regulatory regions

Introduction:

Transposable elements (TEs) are highly repetitive DNA sequences found in most genomes. Variable genome size between related species has been partially attributed to the accumulation of TEs (Michael and Jackson 2013). The maize genome is replete with TEs, having >80% of the ~2500Mb of genomic space being composed of repetitive sequence and 64% annotated as complete TEs (Schnable et al. 2009; Jiao et al. 2016). TEs can be classified into two main orders based on their transposition intermediate, Class I RNA retrotransposons which commonly proliferate through "copy and paste" transposition and Class II DNA transposons that generally move through a "cut and paste" mechanism (Wicker et al. 2007). Barbara McClintock referred to these repetitive sequences as "controlling elements", encompassing their potential to impact and regulate genes (McClintock 1951). Transposition enables these TEs to move throughout the genome potentially influencing functional regions. TEs may insert into coding regions and cause direct influence on gene function, but also may insert into existing regulatory regions or create new regulatory elements resulting in altered gene expression (Lisch2013; Chuong et al., 2017).

One mechanism of TE influence on gene expression is through disruption of regulatory sequences. TEs in the maize genome are dispersed throughout the chromosome including gene-rich regions of chromosome arms (Schnable et al. 2009; Baucom et al. 2009). Due to this interspersion of genes and TEs, many TEs have the potential to influence expression of genes. DNA transposons have been shown to display preferential insertion into genic regions (Dietrich et al., 2002; Liu et al., 2009; Vollbrecht et al., 2010; Springer et al., 2018) while retrotransposons appear to be more present in heterochromatic, gene poor regions of the genome (Bennetzen 2000). In Arabidopsis, MITEs (miniature inverted repeat transposable elements) often insert into the last exon of genes, which may cause more impact than ordinary intron insertions (Guo et al., 2017). A MITE

DNA transposon, mPing in Oryza sativa was found to preferentially insert into the 5' regions of genes (Naito et al. 2009). S-elements in Drosophila melanogaster have been found to insert into the 5' regions of several members of the Hsp70 heat shock gene family (Maside et al. 2002). MITEs and other transposable elements have been hypothesized to play an evolutionary role in altering gene expression through contributing regulatory elements (Wessler et al., 1995; Bennetzen 2000; Lisch 2014).

TEs not only have the potential to disrupt regulatory sequence, but can also introduce novel regulatory elements into new genomic locations (Feschotte 2008; Chuong et al. 2017). TE insertions can also result in changes in the location of regulatory elements relative to nearby genes (Zhao et al. 2018; Lu et al. 2019). It has been shown that TEs can impact gene expression through several examples in maize including *teosinte branched 1* (*tb1*), a gene responsible for the branching in the maize progenitor, teosinte (Studer et al. 2011). The regulatory region of *tb1* is within the intergenic space ~60kb upstream of the gene (Doebley et al. 1997; Clark et al. 2006; Briggs et al. 2007). An essential insertion of a retrotransposon Hopscotch acts as an enhancer of gene expression resulting in the branching differences between maize and teosinte (Studer et al. 2011). Similar examples are observed in other species as well (Zhao et al., 2018; Nishihara et al., 2006; Lowe et al., 2007). The analysis of genes in the human genome have found evidence that TEs may contribute promoters (Jordan et al. 2003) or cis-regulatory regions (Sheffield et al. 2013). The existence of regulatory regions within TEs could represent examples of regulatory elements that have evolved to solely regulate expression of the TE itself as well as examples in which the regulatory elements within the TE have been co-opted to regulate nearby genes (Chuong et al. 2017; Zhao et al. 2018).

The question of how TEs impact the genome has been considered from different perspectives since McClintock first discovered their existence. There are many examples in which detailed analyses of specific QTL have revealed the

importance of TE insertions in creating altered gene expression (Zerjal et al. 2012; Zhang et al. 2012; Yang et al. 2013; Castelletti et al. 2014; Mao et al. 2015). There have been hints that certain families of TEs are associated with genes that exhibit stress-responsive expression (Makarevitch et al. 2015) and that many TEs exhibit dynamic, tissue-specific patterns of expression (Anderson et al. 2019b). There is evidence that a substantial number of accessible chromatin regions are found within TEs (Oka et al. 2017; Zhao et al. 2018; Lu et al. 2019) and in some cases these sequences can provide evidence for regulatory activity (Zhao et al. 2018).

In order to assess the mechanisms by which transposons might influence cis-regulatory elements it is important to have an understanding of putative regulatory elements and transposon variation among genotypes. The availability of genome-wide identification of accessible chromatin regions (ACRs) in B73 (Ricci et al. 2019) and high-quality information on shared and polymorphic TEs (Anderson et al. 2019a) provides new opportunities to address the potential impact of TEs on gene regulation in maize. We characterized hundreds of examples of B73 ACRs that are interrupted by a TE insertion in another genotype and thousands of examples of ACRs that are within annotated TEs. TE insertions into ACRs are associated with chromatin changes to the ACR in addition to the genetic change. Many of these ACRs within TEs show evidence of functional enrichment. Through analyses of putative regulatory regions and TE polymorphisms we can begin to evaluate how TEs may contribute to natural variation for gene expression in maize.

Results

To assess potential impacts of TEs on putative regulatory regions in the maize genome, we used a set of 32,421 previously characterized maize ACRs (accessible chromatin regions) identified using an Assay for Transposase-Accessible Chromatin with sequencing, hereafter referred to as ATAC-seq (Ricci et al. 2019). Roughly similar numbers of ACRs were found within genes

(12,587), proximal regions (within 2kb of genes - 9,183), and distal regions (>2kb from nearest gene - 10,651). Ricci, Lu, Ji et al (2019) documented evidence to support the functional relevance of distal ACRs through enrichment of genetic variants underlying morphological and expression variation (eQTL and GWAS), chromatin-chromatin (HiChIP) interactions, and self-transcribing active regulatory region sequencing (STARR-seq) enhancer activity. We sought to investigate the role that TEs might play in regulating gene expression by disrupting ACRs within the maize genome or in carrying ACRs within TEs (Figure 1A/B). To monitor TE insertions within TE-ACRs, we focused on the set of ACRs identified within the B73 genome (Ricci et al. 2019) and documented the TE insertions in these regions within the W22, Mo17 or PH207 genomes (Figure 1C). The TEs that contain an ACR (>80% of ACR within the TE) were determined by comparing the coordinates of ACRs within the B73 genome with the B73 TE annotations (Figure 1D). The set of TE insertions into ACRs and TEs containing ACRs were further characterized to understand how these changes might influence chromatin states and regulation of nearby genes.

Identification of TE insertions into ACRs

Of the 348 non-redundant instances of TE insertions into B73-defined ACRs, 176 TE insertions were found in Mo17, 82 insertions in PH207 and 158 insertions in W22. To determine the number of TE insertions expected by chance, we used a random set of genomic regions with similar size distribution as the ACRs. We observe significantly (Fisher's exact p-value - 4.286e-07) more TE insertions in ACRs compared to the random regions (Figure S1A). The TEs that inserted were primarily terminal inverted repeat (TIR) DNA transposons with fewer examples of long terminal repeat (LTR) retroelements and Helitrons (Figure 1, Figure S1B). Several TIR elements have been found to be enriched for insertions within accessible chromatin (Kolkman et al., 2005; Liu et al., 2009; Han et al. 2013; Noshay et al. 2019). The insertions into ACRs are highly enriched for members of the DTA and DTM superfamilies (Table S1) of TIR

elements (Figure S1C). The TE insertions located within ACRs tended to represent relatively young TEs based on LTR similarity (Figure S1D).

TE insertions into ACRs can result in altered chromatin

The ACRs represent regions of accessible chromatin and also lack DNA methylation (Ricci et al. 2019). The insertion of a TE in another haplotype could result in not only a genetic change to the DNA sequence, but also to changes in chromatin modifications or accessibility. DNA methylation data was generated for the same tissue type used for ATAC-seq in both B73 and PH207. There are 82 examples of PH207 TE insertions within B73 ACR regions and these were used to investigate the frequency of DNA methylation presence within the region classified as an ACR in B73. Specifically, we assessed the frequency of DNA methylation gains on one (unidirectional), or both (bidirectional) sides of the TE insertion (Figure 2A). In many cases the insertion of a TE within an ACR is not associated with increased methylation of the regions with homology to the B73 ACR (Figure 2B). However, for 37% of the TE insertions within ACRs, there are DNA methylation gains in the haplotype with the TE insertion (Figure 2C). TE insertions that are located within the outer quartiles of the ACR often exhibit methylation gains only on one side of the TE, typically in the region closer to the edge of the ACR (Figure 2D). These analyses were solely focused on TE insertions within the B73 defined boundaries of the ACR. An analysis of 257 additional TE insertions (present in PH207, Mo17, or W22) located within 200bp of the ACR (present in B73) identified 30 additional examples in which a TE insertion near an ACR was associated with DNA methylation gains within the ACR. Together these analyses suggest that a subset of the TE insertions within, or near, ACRs are associated with changes to the DNA methylation state of the region and are likely associated with changes in chromatin accessibility.

Identification of ACRs within TEs

In addition to the potential for TEs to disrupt existing ACRs, they also have the potential to carry sequences that lead to an accessible chromatin state and potentially move these sequences to new genomic locations (Figure 1B). We

focused on characterizing examples of the ACRs that are identified in the B73 genome located within or overlapping annotated TEs. Of the 32,421 identified ACRs in maize, 4,590 have at least a partial overlap with an annotated TE (Table 1). It is worth noting that this is likely an underestimate of the number of true ACRs within TEs as the identification of ACRs relied upon uniquely mapping reads (Ricci et al. 2019). Many TEs are repetitive and have enough similarity to other family members to preclude uniquely mapping reads, which means that the number detected using unique mapping represents only a subset of actual accessible regions within TEs (Figure S5). In both leaf and ear tissue there is no evidence for enrichment of unique mapping reads in ATAC-seq data suggesting the presence of accessible chromatin within repetitive regions (Figure S5A). On a per-TE family basis, in which we could determine the number of reads that map to a family (both multiple mapping and unique mapping reads), there is evidence for some families with substantially more multi-mapping reads (Figure S5B). However, the multi-mapping reads cannot be attributed to a single genomic location and therefore, we focused on the ACRs classified based on unique mapping reads for the remainder of our analyses.

Among the 4,590 TE-ACRs, there are 2,793 examples in which the majority (>80%) of the ACR is located within the TE and another 1,797 that have partial overlap (<80%) (Table 1; Figure S3A). These 1,797 partial overlaps may represent instances in which the ACR within the TE includes some adjacent sequence or may represent instances in which the TE inserted into an existing ACR and the accessible region spreads to encompass a portion of the TE. ACRs within TEs are more common for distal ACRs than for the other types of ACRs, especially for ACRs with majority (>80%) overlap with a TE (Figure S3A). The partial overlaps of ACRs with TEs have a high frequency of TIR elements, while the majority (>80%) overlap TE-ACRs have much higher frequencies of LTR elements (Figure S3A). Given the potential for the partial overlaps to represent instances of TE insertion into or near ACRs, rather than carrying the ACR within

the TE, we focused on the majority (>80%) overlaps for the analyses of ACRs within TEs.

The 2,793 examples of majority TE-ACR overlap mostly (69%) comprise examples of distal ACRs (Figure 1D). Even though only 0.98% of all maize TEs contain an ACR, 19% of the distal ACRs are located within a TE (Table 1). Given an expectation that TEs would not contain accessible chromatin, this represents a large number of unexpected ACRs within TEs. However, if we assume that ACRs are randomly located in genomic sequence then the fact that 19% of distal ACRs are found within TEs is actually substantially fewer than expected (72% of random distal regions with size distribution similar to ACRs overlap a TE) given the amount of sequence attributed to TEs in the maize genome. The distal ACRs were further classified based on the patterns of several chromatin modifications into four groups; K-acetyl enriched, H3K27me3 enriched, transcribed and unmodified (Figure S3B) (Ricci et al. 2019). The TEs containing ACRs are enriched (chi-square p-value < 2.2e-16) for the transcribed class which is characterized by H3K4me3 and H3K36me3 along with acetylation marks and low DNA methylation levels similar to patterns seen in the promoters of expressed genes. This suggests that at least a portion of the ACRs found within TEs may represent promoters for expressed transposable element products. Prior work monitored expression of TEs in a variety of B73 tissues, including pollen and other reproductive tissues (Anderson et al. 2019b). Of the TEs containing an ACR classified as transcribed, 48% show observable expression levels in at least one tissue (Figure S3C). The TEs containing ACRs in the other classes (chromatin marked and unmodified) have lower frequencies of expressed elements but are still expressed more often than non-ACR TEs (Figure S3C).

We investigated the potential that TE-ACRs would be found primarily near highly expressed genes. Using expression data from the same tissue used to perform chromatin accessibility profiling the genes were divided into not expressed (n=13,956) and four expression quartiles (n=6,262 in each quartile)

(Figure S4). As expected, expressed genes were enriched for the presence of ACRs within 5kb of the TSS and highly expressed genes were more likely to have an ACR than low expressed genes (Figure S4a). However, only a small proportion of genes in any group had a TE-ACR within 5kb of the TSS. Highly expressed genes were slightly more likely to have a TE-ACR nearby but in general expressed genes have similar overall numbers of TEs with and without ACRs (Figure S4b). This suggests that some of the TE-ACRs may occur due to proximity to highly expressed genes but also reveals that similar numbers of silent or lowly expressed genes also contain TE-ACRs.

Evidence for potential functional regulatory elements within TEs

Ricci, Lu, Ji et al., 2019 used several approaches to provide evidence for functional impacts of distal ACRs. Focusing on the 10,651 distal (>2kb from nearest gene) ACRs, we sought to determine whether there were differences in the support of functional impact for ACRs within TEs (TE-ACR) compared to ACRs located outside of TEs (nonTE-ACR). The frequency of SNPs is reduced within ACRs and this effect becomes even more pronounced when focusing on the TE-ACRs (Figure 3A). The analysis of the frequency of GWAS- associated SNPs revealed enrichment within both TE-ACRs and nonTE-ACRs (Figure 3B). TE-ACRs also show an enrichment for eQTL, although the level of enrichment is not as strong as observed for nonTE-ACRs (Figure 3C). The difference in the level of eQTL enrichment for TE-ACRs and nonTE-ACRs could be due to the differences in composition among the four chromatin classes of ACRs. The transcribed ACRs generally have lower enrichment than observed for some of the other classes (Figure S6). For ACRs to influence expression they would likely need to interact with nearby gene promoters. HiChIP analysis of chromatin interactions reveal similar enrichment for ACR-genic interactions for both TE and nonTE ACRs (Figure 3D-E). STARR-seq can identify sequences that can provide functional enhancer activity. STARR-seq analysis of maize accessible chromatin fragment activities in maize leaf protoplasts showed similar levels of enrichment for enhancer activity for TE and nonTE ACR sequences (Figure 3F)

Enrichment for certain TE families containing ACRs

TEs are classified into order, superfamily, and family based on transposition mechanism, structural components and sequence similarity. The ACRs that are located within TEs may represent TE family-specific properties in which multiple members of the same family contain an ACR or could represent instances in which the local chromatin neighborhood for a specific TE insertion allows the formation of an ACR. There are 356 (12.7%) of the 2,793 TE-ACRs that are located within single-member TE families, which is much greater than the overall frequency (1.5%) of single copy TEs in the genome. Among the remaining 2,437 TE-ACRs that are within multi-member TE families, 557 are only in one of the TEs in the family containing an ACR. This suggests that the majority of TE-ACRs are not a reproducible feature of the family members. A caveat to these results is the repetitive sequences which would not have been captured through the unique mapping ATAC-seq analysis and therefore additional members of a family may contain accessible chromatin regions (Figure S5B).

There are examples of TE-ACRs that are found in multiple members of a TE family. There are 112 TE families with at least two members with an ACR. There are only 10 of these families (with at least 3 elements) in which >30% of the elements have an ACR (Figure S7A). These examples of TE families with multiple members with ACRs were identified based on utilization of unique mapping reads. It is quite possible that additional members of these families may contain ACRs that were not identified because they are in regions that are highly similar in multiple TEs and therefore are multi-mapping. Two families in particular, RLX00813 and RLX01441, were found to display increased coverage when multi-mapping was allowed (Figure S7B).

ACRs within TEs show variable DNA methylation patterns among genotypes

In general, TEs are considered to have quite high levels of DNA methylation, but ACRs typically lack DNA methylation (Oka et al. 2017; Lu et al. 2019; Ricci

et al. 2019). The presence of ACRs within TEs led us to investigate the DNA methylation level of these sequences. We found that while TEs containing an ACR show quite high levels of DNA methylation throughout most of the TE, the ACR section is essentially unmethylated (Figure 4A-B). Visual inspection of several examples reveals that the ACR region represents a small window of unmethylated DNA within the largely methylated TE (Figure 4C-D).

We hypothesized that the presence of an unmethylated region within a TE might be somewhat unstable and could be subject to changes in DNA methylation state among different haplotypes at a higher frequency than ACRs not located within TEs. An analysis was performed using a set of B73 ACRs that have a matching sequence at a syntenic location in PH207, Mo17, or W22 and have DNA methylation data available for both genotypes. These include ACRs within TEs that are present in both genomes and ACRs that are present in non-TE sequence (nonTE ACRs). While less than 3% of the nonTE ACRs exhibit gains of CG methylation across each of the genotypes, there are over 12% of the ACRs that are located within TEs that exhibit high levels of CG methylation (Figure 5A). Visual inspection of several loci suggests gains of both CG and CHG methylation over the full ACR sequence in these examples (Figure 5B-C). These observations suggest that ACRs within TEs may exhibit less stability among genotypes than ACRs in nonTE regions of genomes.

TE presence association with gene expression

Polymorphic TEs that interrupt an ACR or create novel ACRs in some haplotypes have the potential to influence the expression of nearby genes. To assess the potential for these polymorphic TE-ACR interactions to influence gene expression, we sought to associate the presence/absence of TEs with the changes in relative expression levels for nearby genes in panels of diverse germplasm. De novo assembled genome sequences of B73, Mo17, PH207 and W22 were used to generate de novo TE annotations in these four genomes (Anderson et al. 2019a). The presence or absence of these TEs was assessed in a larger (>500 inbred lines)

panel of diverse maize lines using alignments of whole-genome shotgun sequencing reads to the TE-flanking sequence junctions (O'Connor et al. 2020). This approach provides robust assignments of presence or absence for many genotypes but in some cases there is not clear evidence and the TE status is classified as ambiguous in that genotype. The TE polymorphism information was used to investigate variation in gene expression in several RNA-seq datasets (Hirsch et al. 2014; Kremling et al. 2018; Mazaheri et al. 2019). Each of these datasets included samples from a panel of genotypes that were collected at similar tissue stages.

Each polymorphic TE that disrupts a B73 ACR or contains an ACR in B73 was assigned based on HiChIP interactions or proximity to the nearest gene. TE-gene pairs where the gene is present completely within an annotated TE were disregarded for this analysis. We then assessed the difference in expression for genotypes with or without the TE insertion across the two datasets incorporating 284 genotypes and 8 tissues. (Table 2; Figure S9) allowing separate tests of potential associations between TE polymorphisms and expression level in multiple tissues. We initially focused on the set of 377 TE insertions into an ACR, which we hypothesized may result in reduced expression for the nearby gene. The majority of these TE insertions into ACRs have limited associations with the expression of nearby genes. There are 21 instances (5.6% of all TE-gene pairs) in which we found a significant (q-value <0.05 and >2-fold-change) change in expression for the nearby gene (Table 2). These include 9 genes in which higher expression was observed for the haplotype containing the TE insertion, and 12 examples of lower expression when the TE is present. In 10 of the 21 significant associations, we found a significant association between the presence of the TE and expression levels in multiple tissues. In addition to the genes with significant associations, we also noticed that there is an apparent excess of many 'outlier' expression states for which the genotype with (or without the TE) has a >30-fold change in expression but there is limited statistical significance because one of the haplotypes is rare (Figure S9A). To determine

if there is a significant excess of these outliers, we performed separate permutation tests in which the genotype-expression or genotype-TE presence classifications were randomized. These were separately performed for each of the expression datasets and were used to determine the number of significant or outlier expression changes expected by chance within this data structure (Figure 6A). The TE insertions into ACRs consistently exhibit more outliers than expected by chance with reduced expression of the haplotype with the TE present for each of the expression datasets (Figure 6A).

We next assessed the 2,182 polymorphic insertions of TEs containing ACRs near genes which were hypothesized to have positive influences on the expression of the nearby gene. There were 190 significant associations (8.7% of all tested TE-gene pairs) and 81% of these significant associations exhibit higher expression for the nearby gene (Figure S9B, Table 2). Many (49%) of the significant positive associations between the presence of the TE and the expression of the nearby gene were identified in multiple tissues while fewer (18%) of the negative associations were identified in multiple tissues. Figure 6C-D shows two examples of a TE located near a maize gene with significant positive associations with expression in multiple tissues. In both of these examples there are HiChIP interactions between the ACR within the TE and the nearby gene based on data from Ricci, Lu, Ji et al (2019). The permutations tests identify very few significant associations (Figure 6B). The analysis of rare outlier expression states also reveals an excess of positive associations in which the haplotype containing the TE exhibits a higher expression level (Figure 6B). To further support the cis-regulatory variation observed at the examples of significant associations between presence of TE-ACRs and expression of nearby genes we evaluated allelic bias for expression in F1 hybrids. Prior work had generated allele-specific expression for 23 tissues in the B73 x Mo17 F1 hybrid (Zhou et al., 2019). There are 26 polymorphic TE-ACR insertions in B73-Mo17 with significant associations with expression and allele-specific data available. When we investigate tissue types most closely related to the tissue with significant

associations we find significant allelic expression bias for 19 of these 26 genes in the predicted direction (Figure S10). Most of the 7 genes without significant allelic bias still exhibit a bias in the expected direction but did not contain sufficient sequencing depth to provide evidence for significant effects. This further confirms the presence of cis-regulatory variation for these loci.

Discussion

Many eukaryotic genomes show evidence for both recent amplification of transposable elements as well as turnover of elements through deletions (Bennetzen and Kellogg 1997). Insertions of transposons into genes or regulatory elements can lead to loss-of-function mutations which are presumed to be primarily deleterious. However, there is growing evidence that TEs may also contribute to re-wiring of transcription of nearby genes (Weil and Martienssen 2008; Feschotte 2008; Lisch 2013; Chuong et al. 2017). Transposon insertions that affect expression of a nearby gene are the molecular basis for allelic variation at several loci important for maize domestication and improvement (Studer et al. 2011; Yang et al. 2013; Castelletti et al. 2014). There are also examples in maize and other species in which transposon insertions may influence regulatory influences on nearby genes (Jiang et al. 2004; Cavrak et al. 2014; Makarevitch et al. 2015; Zhao et al. 2018). While specific examples have been identified, the genome-wide frequency for these TE influences has not been characterized. Advances in our knowledge of genome-wide TE polymorphisms (Stitzer et al.; Anderson et al. 2019a) as well as the identification of proximal and distal putative cis-regulatory elements (Oka et al. 2017; Zhao et al. 2018; Ricci et al. 2019) provided an opportunity to assess the mechanisms and frequency by which TEs may create regulatory variation

In this study, we focused on two potential ways in which TEs might influence the expression of nearby genes; the disruption of regulatory regions and the introduction of novel sequences that may act as regulatory sequences. Insertions into regions of accessible chromatin might be expected to often result in reduced

expression of nearby genes or altered patterns of expression. In contrast, TEs that contain accessible chromatin regions may be mobile enhancers that affect expression of both the TE promoter as well as nearby gene promoters. Several studies have found that putative enhancers can be found within transposable elements in the maize genome (Oka et al. 2017; Zhao et al. 2018). We were interested in assessing how frequently the polymorphic insertions could be associated with variable expression for nearby genes to understand the potential for TE polymorphism to generate regulatory diversity. It is worth highlighting the fact that truly assessing the potential for TEs to influence regulation in natural populations may be complicated by the potential fitness consequences of polymorphic TE insertions. If a TE insertion results in significant deleterious or beneficial consequences the allele will likely be a target of selection. Recent studies have found that there are likely many examples of rare deleterious expression states in domesticated maize populations (Kremling et al. 2018) and therefore we monitored both common and rare expression states associated with TE polymorphisms.

Potential for TEs to reshape chromatin and the epigenome

Active transposition of TEs results in genetic changes including disruption of genes or regulatory elements as well as potential genomic instability due to chromosome breaks or illegitimate recombination. To limit these deleterious events, most genomes have evolved mechanisms to restrict active transposition, including epigenetic silencing through chromatin modifications such as DNA methylation (Hollister and Gaut 2009; Lisch 2013; Springer et al. 2016). This results in highly methylated TEs in plant genomes (Niederhuth et al. 2016) and has been observed to spread outside of the TE sequence to surrounding DNA sequences in some cases (Wyler et al., 2020; Choi and Puruggana 2018; Eichten et al. 2012; Noshay et al. 2019). As TEs insert into putative regulatory regions, the question becomes not only how the presence of new DNA sequence impacts this region but also the potential for alteration of chromatin patterns. The TE insertion into regions of accessible chromatin can potentially result in loss of

accessibility and gains of DNA methylation for the flanking sequences. We observe many examples of TE insertions into accessible chromatin regions for which the regions immediately flanking the TE remain unmethylated and potentially accessible. In some cases, the insertion of a TE within a larger accessible chromatin region results in two smaller accessible chromatin regions on either side of the TE. Often these regions have partial overlap with the edges of the TE. However, there are a subset of examples of TE insertions into accessible regions where the previously accessible and unmethylated regions exhibit high levels of methylation on one or both sides of the TE insertion in the TE-present genotype.

TEs that introduce novel accessible chromatin regions have the challenge of maintaining an unmethylated accessible chromatin region within a highly targeted and condensed repetitive sequence. Even in the TEs that contain an accessible chromatin region, we find that the remainder of the TE is highly methylated. When assessed across three additional genotypes, the methylation state of these accessible chromatin regions was more variable than other unmethylated regions that were outside of TEs. This may suggest that the presence of a TE containing a putative regulatory element in the B73 genome may not predict the presence of an active regulatory element in other genotypes. These would result in the potential for facultative epialleles (Richards 2006; Springer and Schmitz 2017) in which some haplotypes with the TE contain an active regulatory element while others would have a silencer element. This would complicate our ability to make associations between the genetic presence/absence of the TE and the expression level of nearby genes. In our analyses, we made the assumption that when the TE is present the accessible, unmethylated region will be conserved. However, epigenetic polymorphisms would significantly reduce our power. Indeed, careful examination of some examples such as those in figures 6C and D reveal that even though the TE presence is often associated with higher expression for the nearby genes there are some haplotypes that contain the TE but do not show high expression for the

nearby gene. These may reflect epigenetic silencing of the regulatory element within these TEs. Alternatively, this could reflect potential variation in trans-acting factors.

TE influences on regulatory variation for genes

There are massive numbers of polymorphic TE insertions between any two maize genotypes (Wang and Dooner 2006; Springer et al. 2018; Sun et al. 2018; Anderson et al. 2019a). The majority of these polymorphisms likely have little or no impact on gene products or gene expression and are essentially neutral polymorphisms. However, if even a small portion influences gene expression, this could account for a major source of regulatory variation. In this study, we have used chromatin accessibility profiling to narrow the set of TE polymorphisms that might result in altered expression for nearby genes. Specifically, we focused on two classes of polymorphisms that could be assessed based on high quality chromatin accessibility data for the B73 genome (Ricci et al. 2019). The presence of an accessible chromatin region within a TE in B73 enables us to investigate whether the presence of this TE in other maize genotypes is associated with high, or lower, expression of the nearby gene. Alternatively, the presence of an ACR in B73 with a polymorphic TE insertion in PH207, Mo17, or W22 allows for an understanding of how the interruption of an ACR may influence gene expression.

Even in this focused set of TE polymorphisms we find that most of the TE polymorphisms are not significantly associated with altered expression of nearby genes in the tissues we monitored. A majority of genes were found to have little to no change in expression level relative to TE presence/absence (80% of TE-ACRs and 87% of TE insertions into ACRs). This could suggest that these TE-ACRs do not influence expression of the nearby gene. However, it is also possible that in some cases we have not examined the right tissue or growth condition, or that epigenetic instability of the ACR within TEs might complicate our ability to make a genetic association as described above. While the majority

of TE polymorphisms were not significantly associated with expression for nearby genes, there are 21 examples of TE insertions into ACRs and 190 examples of TE containing ACRs that are significantly associated with the expression of nearby genes. The lack of strong effects for TE insertions into ACRs was somewhat surprising. In some cases, the TE insertions into ACRs may result in dividing a single ACR into two regions separated by the TE. This would predict that there would be instances in the B73 genome in which there are two nearby ACRs that are separated by a TE and the insertion did not necessarily disrupt the functionality of the regulatory region. Interestingly, the examples of TE containing ACRs that are significantly associated with expression are heavily biased towards examples in which the nearby gene is higher expressed. This suggests the TE is providing an enhancer that increases gene expression. In addition to the significant associations, there are also many other examples in which there is substantial variation in expression levels for haplotypes with and without the TE, but which lack any statistical significance (outliers). These likely represent examples in which the haplotype with (or without) the TE is rare and only present in one or two genotypes. This might be expected in situations in which TE insertions influence expression resulting in substantial deleterious effects. These outliers are enriched for lower expression of the nearby gene for TE insertions into ACRs but higher expression for the nearby gene for TEs containing ACRs.

A key question we wrestled with in this study, is whether the presence of an ACR within a TE was a property of certain TE families. Given the sequence conservation within TE families, we might predict that the presence of a regulatory element would be conserved in many members of the same TE family. Searching for this consistency is complicated by the focus on uniquely mapping reads. Indeed, we have likely greatly underestimated the number of ACRs within TEs (Figure S5). In many cases, we would only find an ACR in one member of a multi-TE family. These might suggest that the ability to form an accessible region is attributed to both the genetic sequence of the TE as well as local

chromatin context. We do find examples of TE families in which there are multiple members with an ACR but even in these families there are other members that lack the ACR (Figure S7-7). In this analysis we do not find strong evidence for TE families in which a common regulatory element is present and accessible for many elements of the same family. This highlights the role for both the DNA sequence of TEs as well as the chromatin landscape of these TEs.

Identification of accessible chromatin regions across the genome has enabled us to narrow in on the ~1% of the genome with potential regulatory function (Rodgers-Melnick et al. 2016; Oka et al. 2017; Zhao et al. 2018; Ricci et al. 2019). By assessing how TE variation could contribute to polymorphisms for these accessible regions we have characterized the potential for TEs to disrupt ACRs or contribute novel ACRs to genes. We assessed both the chromatin and regulatory consequences of these polymorphisms. We find evidence that a subset of TEs containing ACRs are likely providing enhancers to nearby genes. There was little evidence for widespread consequences of insertions of TEs into ACRs. However, many of the TE polymorphisms that strongly influence gene expression might represent rare deleterious alleles. This analysis highlights the potential for TEs to influence gene expression by creating novel expression patterns rather than simply disrupting existing information.

Methods

Data Availability:

In this study we utilize datasets that are available through the following accessions: SRX4727413, SRR8738272, SRR8740852, and BioProject PRJNA661271.

Annotation of Genes and TEs:

Whole genome assemblies for B73 (Zm00001d) (Jiao et al. 2016), W22 (Zm00004b) (Springer et al. 2018), Mo17 (Zm00014a) (Sun et al. 2018), and PH207 (Zm00008a) (Hirsch et al. 2016) were used for genome-wide analyses.

All analyses were done on assemblies of chromosomes 1-10 (the canonical maize chromosomes) while all un-placed scaffolds were disregarded due to the inability to compare these regions across genotypes. Filtered structural TE annotations (Stitzer et al.; Anderson et al. 2019a) were used.

Polymorphic TEs

Shared and non-shared TEs across genotypes were defined previously (Anderson et al. 2019a). Briefly, identification of shared and non-shared elements was determined through pairwise comparison between four maize inbred lines (B73, W22, PH207, and Mo17). Cross-genotype gene keys were generated using scrips available at https://github.com/SNAnderson/maizeTE_variation/gene-key_pipeline. Gene syntelogs were defined by a multi-approach method described in Anderson et al. (2019) combining SynMap, Nucmer, and OrthoFinder. Search windows were defined by the closest, non- overlapping genes to the query TE with a syntelog in the genome being assessed. For comparison, 400bp flanking tags were extracted for each annotated TE in the genome (for each genome assessed) centered at the start and end coordinates. These flank tags were mapped to the other genomes with use of BWA-MEM (Li and Durbin 2009) in paired-end mode. Further characterization was performed on those elements with tags mapped completely within the search window. Non-shared site-defined TEs were defined by alignment of only the outer 200bp of the flank tags where the distance between tags was less than twice the TSD length for the superfamily. This resulted in a total of 69,292 non- shared site-defined elements across all pairwise comparisons used for analyses (Anderson et al. 2019a).

A total of 509,629 non-redundant TEs defined in at least one of B73, Mo17, PH207 or W22 structural TE annotations were assigned as present or absent in 509 of the WiDiv inbred genotypes (O'Connor et al. 2020; Hansey et al. 2011). Methods for classification of present/absence transposable elements are described in O'Connor et al. 2020 (BioRxiv 10.1101/2020.09.25.314401).

Briefly, two points of reference, 10 bp over left and right inner edges of a TE, were used to determine TE status in a particular genotype. TEs with a coverage >= 8 across both inner edges were classified as present while TEs with coverage < 7 across both inner edges were classified as absent. All other TEs were classified as ambiguous. All TEs defined as present and absent in at least one other genotype were maintained for downstream analyses (PAV calls across the 509 inbred lines for each TE can be found in the DRUM database: http://hdl.handle.net/11299/216935). Data presented in O'Connor et al. (2020) only uses a subset of this TE list based on a frequency threshold of genotypes with an ambiguous classification. Sequencing data (with >20X coverage) for each of the 509 inbred maize genotypes are available at SRA (BioProject PRJNA661271).

Methylation data:

In this study we utilized previously generated WGBS data for B73 seedling shoot, PH207 seedling shoot, Mo17 seedling leaf and W22 seedling leaf. Trim_glore (Martin 2011) was used to trim adapter sequences and read quality was assessed with the default parameters and paired-end reads mode. Reads that passed quality control were aligned to the B73v4 genome (non-B73 genotypes were also aligned to their corresponding genome assemblies). Alignments were conducted using BSMAP-2.90(Xi and Li 2009), allowing up to 5 mismatches and a quality threshold of 20 (-v 5 -q 20). Duplicate reads were detected and removed using picard-tools-1.102 ("Picard Tools - By Broad Institute") and SAMtools (Li et al. 2009). Conversion rate was determined using the reads mapped to the unmethylated chloroplast genome. The resulting alignment file, merged for all samples with the same tissue and genotype, was then used to determine methylation level for each cytosine using BSMAP tools. Methylation ratios for 100bp non-overlapping sliding windows across the B73v4 genome in all three sequence contexts (CG, CHG, and CHH) were calculated

(#C/(#C+#T)). Each 100bp window was categorized as methylated (>=40%), intermediate (20-40%), or unmethylated (<=20%) based on the CHG methylation level.

ATAC-seq data:

In this study we utilized previously generated seedling shoot ATAC-seq data for B73 (Ricci et al. 2019). Raw reads were trimmed with Trimmomatic v0.33. Reads were trimmed for NexteraPE with a maximum of two seed mismatches, palindrome clip threshold of 30, and simple clip threshold of 10. Reads shorter than 30 bp were discarded. Trimmed reads were aligned to the Zea mays AGPv4 reference genome 44 using Bowtie v1.1.147 with the following parameters: "bowtie -X 1000 -m 1 -v 2 --best –strata". Aligned reads were sorted using SAMtools v1.3.1 and clonal duplicates were removed using Picard version v2.16.0 (http://broadinstitute.github.io/picard/).

Identification of accessible chromatin regions (ACRs):

MACS2 was used to define accessible chromatin regions (ACRs) with the "--keep-dup all" function and with ATAC-seq input samples (Tn5 transposition into naked gDNA) as a control. The ACRs identified by MACS2 were further filtered using the following steps: 1) peaks were split into 50 bp windows with 25 bp steps; 2) to quantify the accessibility of each window, the Tn5 integration frequency in each window was calculated and normalized with the average integration frequency across the whole genome to generate an enrichment fold value; 3) windows with enrichment fold values passing a cutoff (25-fold) were merged together by allowing 150 bp gaps; 4) to remove possible false positive regions, small regions with only one window were filtered for lengths > 50 bp. The sites within ACRs with the highest Tn5 integration frequencies were defined as ACR "summits".

For the functional analysis of SNP, HiChIP, STARR-seq and eQTL data we utilized the same methods as described in Ricci, Lu, Ji et al., 2019. The

difference lies in the subset of data that was used to focus on TE ACRs versus non-TE ACRs opposed to all distal ACRs in the genome.

Determination of TE-ACR overlap:

TE-ACRs were defined by an overlap of B73 ACR coordinates with the structural TE annotation coordinates. Each ACR was assigned to a single TE using bedtools closest based on the disjoined TE coordinates file. For those with a partial overlap of multiple TEs the ACR was assigned to the TE with the greatest overlap. Complete overlaps were defined by
>80% of the ACR length overlapping a TE.

Identifying TE-insertions into ACRs:

Site-defined TE polymorphisms with the TE present in Mo17, W22, and/or PH207 and absent in B73 were utilized to identify TE insertions into ACRs. Bedtools intersect was run with all defined B73 ACRs and the site-defined insertions, using the B73 insertion site coordinates. Any site-defined TE in Mo17, PH207, and/or W22 that had an insertion site within the coordinate range of a B73 ACR was characterized as a TE-insertion into an ACR for further analyses.

A set of control regions were generated as a genome-wide proxy for potential accessible regions. The genome was subset to “mappable” sequence determined by WGBS read coverage and used as the input to bedtools shuffle along with the identified ACRs. Output contains the same number of regions with the same lengths as the ACR input file randomly placed across the mappable genome. These regions are used as a control for the frequency of insertions into accessible regions.

Analysis of methylation at TE insertion sites:

Methylation for each TE insertion was defined for the TE present genotype (Mo17, PH207, or W22) and the TE absent genotype (B73). Changes in methylation were identified by comparing 100bp bin CG methylation of the

ACR in B73 to CG methylation levels flanking the insertion site in the genotype present for the TE. The position of the insertion was determined by its location in the ACR by quartiles with the 1st and 4th quartile being insertions at the edge of the ACR and the 2nd and 3rd quartiles defined as insertions into the middle of the ACR.

Analysis of Methylation at ACRs across genotypes:

Gene anchor files have been one to one gene syntelogs pairwise between B73, Mo17, PH207, and W22. Gene key files are available at https://github.com/SNAnderson/maizeTE_variation and were filtered to only one-to-one gene matches. Bedtools closest upstream and downstream, ignoring overlaps, was run for each B73 ACR relative to gene anchor files between B73 and PH207, W22, and Mo17. The search window was defined by the closest upstream and downstream non-overlapping genes in the query genome on either side of the ACR sequence that has a unique syntelog in the target genome. BLAST was run for each B73 ACR sequence to PH207, W22, and Mo17 to identify sequence similarity in the search window for the corresponding genotype. The sequence coordinates were identified and bedtools overlap was run against the 100bp WGBS data for that genotype. The methylation state of the B73 ACR was compared to the methylation levels of the matching sequence in PH207, W22, and Mo17 (based on WGBS data aligned to the corresponding genome assembly). The ACR was characterized as methylated if the average level of CHG methylation was greater than 40% and unmethylated if the average level of CHG methylation was less than 20%. A change in methylated was identified by an ACR characterized as unmethylated in B73 having a methylated state in another genotype.

Gene expression analyses:

RNAseq datasets Hirsch et al. (Hirsch et al. 2014) and Kremling et al. (Kremling et al. 2018) were used to assess expression levels across 284 genotypes and 8 tissues (Table 2). To assess gene expression variation, the

closest gene to each TE was determined in B73 and the expression of that gene was associated with the presence or absence of the TE in each of the 284 genotypes. Each element containing an ACR or inserting into an ACR was assigned to the closest B73v4 annotated gene (in either direction) using bedtools closest. Only one assignment was given for each TE and any TE annotated as containing the full sequence of a gene was removed from the analysis. For those with distal ACRs, HiChIP data was used to assign the gene if an interaction was identified (Table S2/S3). TE presence impact was determined for each TE-gene pair by averaging the expression values for TE- present genotypes and TE-absent genotypes and the log2(present/absent) value was calculated. To account for biases in the number of genotypes with each TE as present or absent a t-test was performed to determine the p-value for each gene in each tissue.

Tables:

Table 1: B73 ACRs majority overlapping (>80%) or partially overlapping (<80%) annotated TEs

	Genic	Proximal	Distal
Total	12587	9,183	10,651
LTR	138 (93)	130 (94)	1428 (225)
TIR	25 (382)	72 (387)	63 (376)
Helitron	301 (90)	203 (74)	433 (76)
Total TE	464 (565)	405 (555)	1924 (677)

* values in () represent partial overlaps (< 80%)

Table2: RNA-seq and TE PAV dataset summaries

Dataset	# Tissues	Genotypes w/ TE calls	TE-Insertion (N=377)		TE-ACR (N=2182)	
			Significant (+/-)	Outliers (+/-)	Significant (+/-)	Outliers (+/-)
Kremling et al. (2018)	GRoot	91	2/ 1	16 / 17	51 / 2	214 / 59
	GShoot	91	3 / 10	0 / 27	55 / 4	204 / 54
	Kern	84	4/ 3	15 / 23	67 / 2	240 / 60
	L3Base	87	2/ 4	19 / 25	54 / 4	197 / 65
	L3Tip	86	5/ 1	19 / 22	44 / 6	281 / 60
	LMAD	54	3/ 0	17 / 27	30 / 8	265 / 86
	LMAN	94	0/ 3	14 / 32	52 / 11	256 / 73
Hirsch et al. (2014)	Seedling	230	1/ 2	2/ 7	57 / 14	105 / 22
Non-redundant sum	All of the Above	259	9 / 12	57 / 86	153 / 37	667 / 295

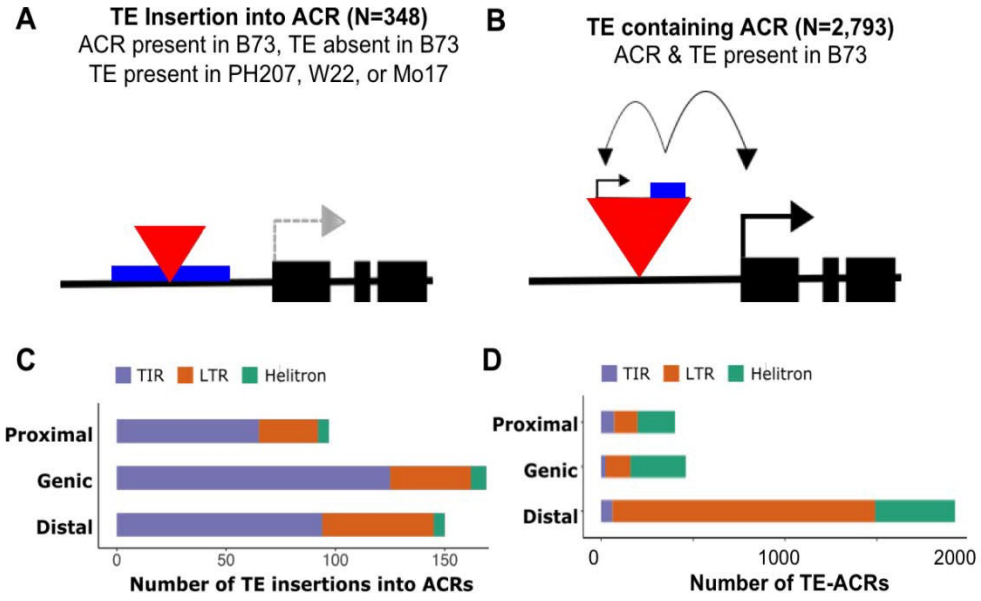

Figure 1: An overlap of TEs and accessible chromatin regions (ACRs). Schematic representation of the identified ACRs (blue) in the B73 maize inbred line and their interaction with TEs (red) and the potential impact on nearby genes. A) B73 ACRs that have a site-defined TE insertion in Ph207, Mo17 or W22. B) B73 ACRs that are found within B73 TE sequence. C) The number of TE insertions (as shown in A) in PH207, Mo17, or W22 into each ACR category (characterized by their position relative to annotated genes as genic, proximal, or distal) of ACR based on site-defined insertion sites in B73. Colors represent TE order. D) Number of TE-ACRs (as shown in B) by location relative to genes and TE order.

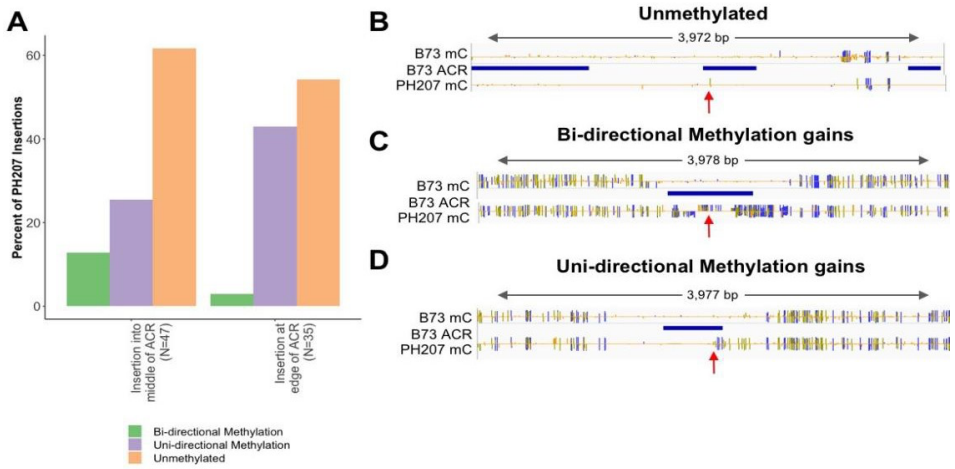

Figure 2: Methylation changes due to TE insertions in PH207. A) For every PH207 site-defined TE insertion into a B73 ACR the PH207 methylation status is defined as unmethylated (region remains unmethylated just as it was in B73), uni-directional methylation (methylation gain on one side of the insertion site), or bi-directional methylation (methylation gain on both sides of the insertion site). Insertions are broken into those that insert into the middle of an ACR (quartile 2 or 3) or those that insert into the edge of an ACR (quartile 1 or 4). WGBS data for B73 and PH207 were aligned to the B73 genome to visualize. IGV views display methylation level tracks (blue is CG, green is CHG, yellow is CHH), ACR region tracks, and TE insertion sites indicated by red arrows. These are shown for each methylation status; B) unmethylated, C) bi-directional methylation, D) unidirectional methylation.

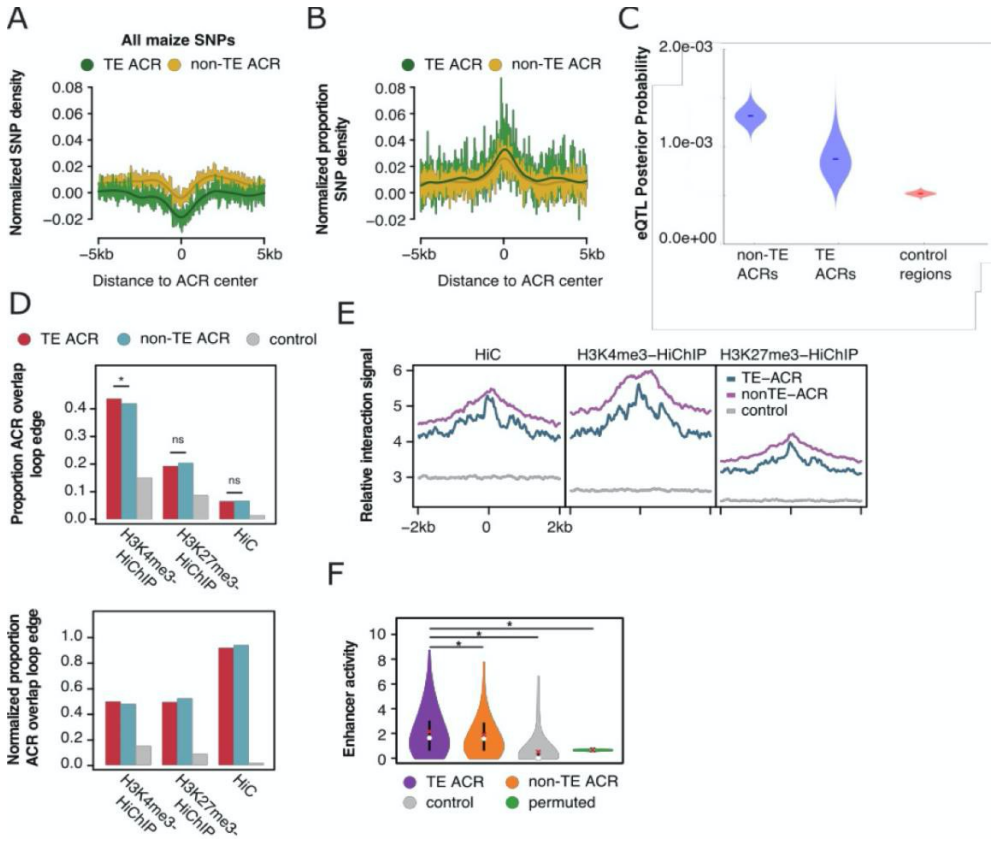

Figure 3: Functional differences between TE and non-TE accessible chromatin regions among distal ACRs. A) Normalized (control) SNP density among maize inbred lines averaged across 10kb regions centered on TE and non-TE ACRs. B) Proportion of GWAS hits (out of all maize SNPs) normalized by control enriched within 10kb windows centered on TE and non-TE dACRs. C) eQTL posterior probability for TE and non-TE ACRs compared to control regions. D) Contrasts between the proportions of dACRs overlapping an I-G loop between TE-ACRs and non-TE ACRs. Chi-square, *P-value <0.05. E) Relative enrichment of chromatin interaction tags across 4kb windows centered on TE ACRs and non-TE ACRs across the three types of chromatin loops. F) Distribution of enhancer activities for dACRs split by the presence/absence of TEs, control regions (n=4,406) and the means of a permutation (10,000x). Statistical differences between TE and non-TE ACRs were evaluated with Mann-Whitney rank sum test. Statistical differences between distribution means and permuted regions were estimated as empirical P-values. ns, not significant; *P < 0.05

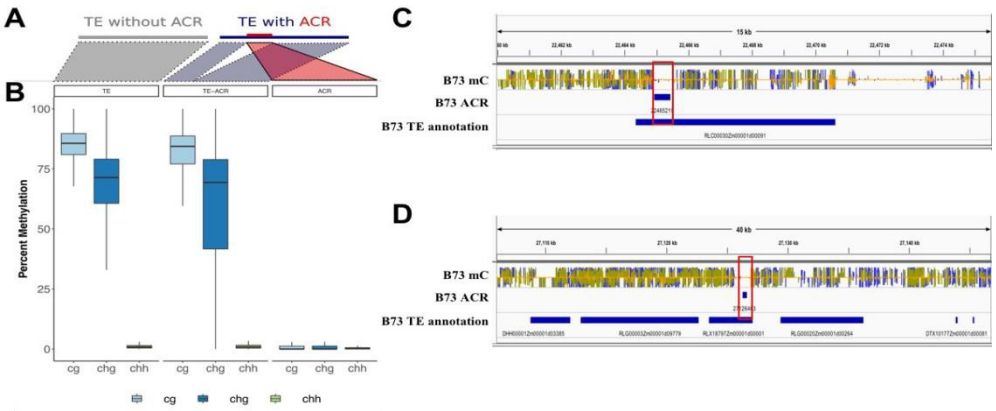

Figure 4: TE-ACR methylation patterns. A) Schematic representation of a TE without an ACR (grey) and a TE containing an ACR (blue) with the ACR sequence shown in red. B) Methylation levels of TEs without ACRs, TEs with an ACR (excluding ACR bins), and ACRs showing the trend that TEs maintain similar levels of high CG and CHG methylation with and without an ACR but the ~300bp region of an ACR is unmethylated. C/D) IGV view of TE with an ACR and the methylation levels (CG blue, CHG green, CHH yellow) over a majority of the TE and absence over the ACR.

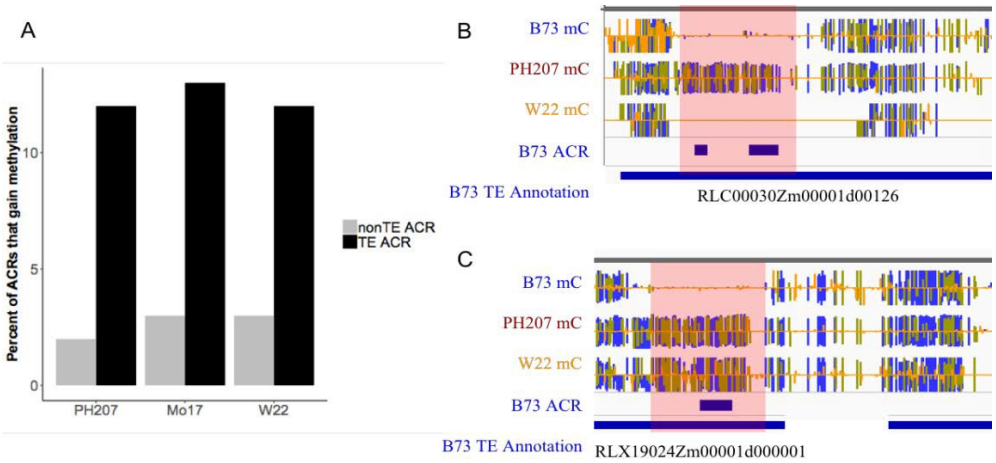

Figure 5: Unmethylated (open chromatin) regions in TEs are less stable than nonTE open chromatin regions. A) Percent of ACRs that gain methylation in PH207, Mo17, or W22 for non-TE ACRs (grey) and TE ACRs (black). B/C) IGV view of B73 TE annotation with unmethylated ACR in B73 and the same region as methylated in PH207 and/or W22. Methylation tracks show CG methylation in blue, CHG methylation in green, and CHH methylation in yellow.

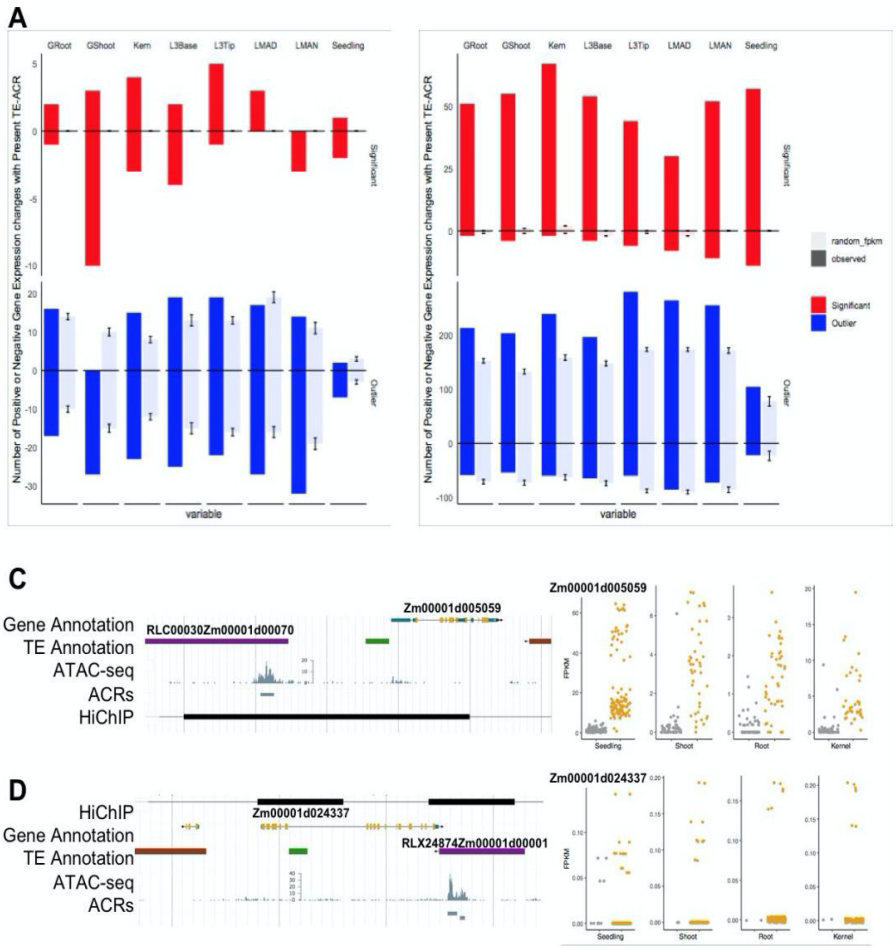

Figure 6: TE PAV association with gene expression. A) Number of TE-Insertions that result in significant (red) or outlier (blue) expression changes of nearby genes by tissue for observed and randomized genotype or randomized RNA-seq controls shown by shading. B) Number of TE-ACRs resulting in significant or outlier expression changes. C/D) Examples of significant gene expression changes associated with TE presence. Left: Genome browser view of the TE, Gene, and ACR. Right: Dot plot of gene expression for genotypes present (yellow) or absent (grey) for seedling, shoot, root, and kernel corresponding to the TE-Gene pair.

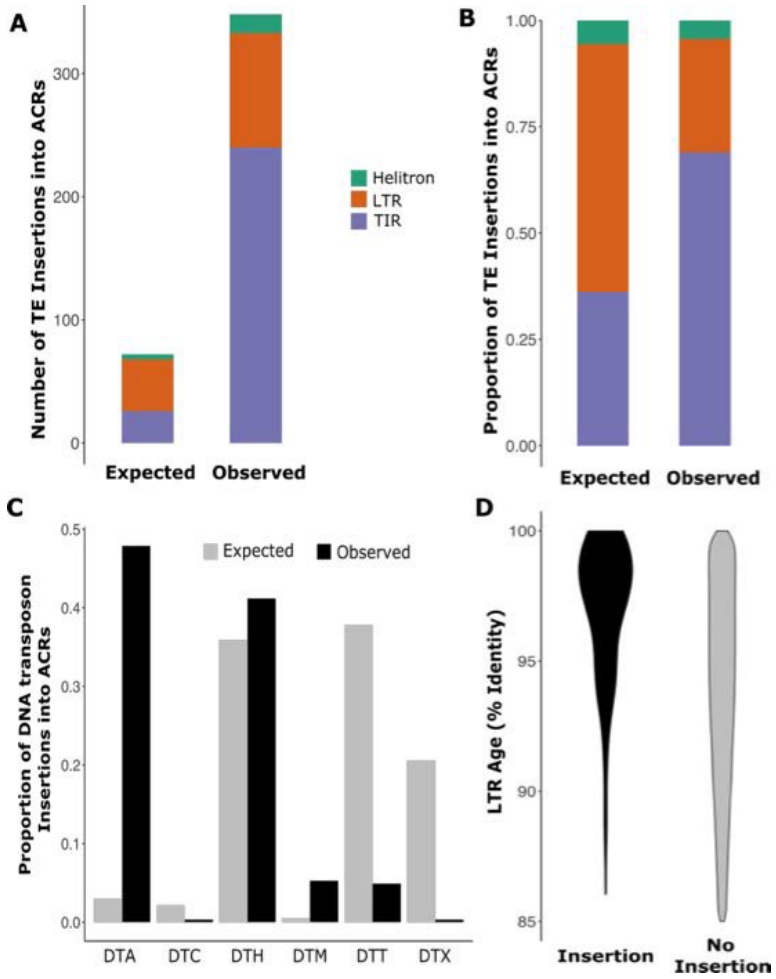

Figure S1: TE insertions by superfamily. A) Raw number of TE insertion into ACRs identified (observed) and a control set of random regions of the same size (expected). (B) The proportion of TE insertions into ACRs that are TIRs (purple), LTRs (orange), or Helitrons (green) relative to that expected by chance based on randomized regions of the same size. C) Proportion of DNA transposons that belong to each superfamily for observed (black) or expected based on randomized regions (grey) insertions into ACRs. D) LTR insertions (black) are younger on average than all LTRs in the genome (grey). LTR age is determined by percent identity of the LTR sequences (high % identity represents younger TEs).

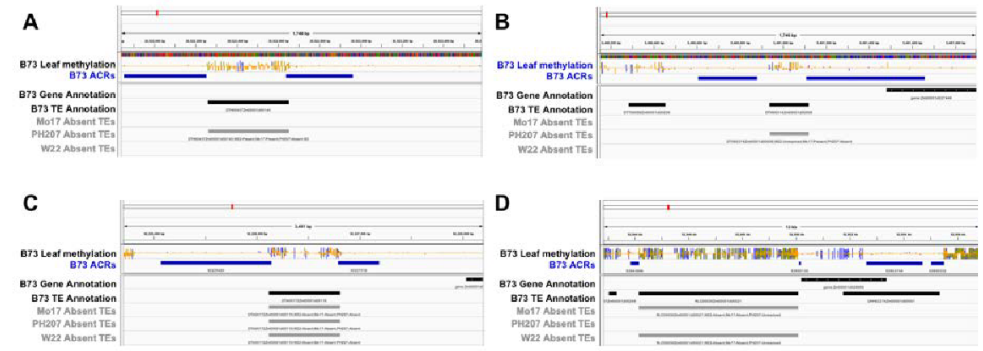

Figure S2: TE insertions split ACRs. TE insertions into B73 ACRs may result in unmethylated regions on either side of the TE in other genotypes suggesting a TE may split accessible chromatin regions. IGV views display tracks with B73 WGBS methylation (CG blue, CHG green, CHH yellow), B73 ACRs, B73 gene annotations and B73 TE annotations. Each panel identifies a case where a B73 TE is flanked by ACR fragments and the TE is polymorphic in another genotype. A) distal B73 TE absent in PH207, B) proximal B73 TE absent in PH207, C) proximal B73 TE absent in PH207, W22, and Mo17, and D) proximal B73 TE absent in Mo17 and W22.

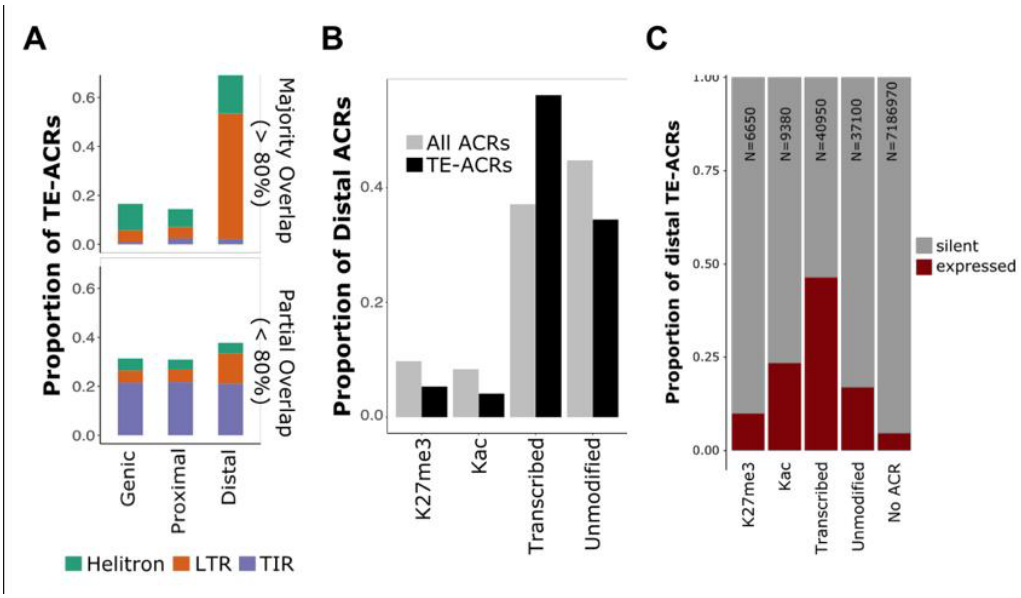

Figure S3: TE-ACR characterization. A) Proportion of all ACRs in each location category that overlap a TE, majority (>80%) or partial (<80%). Color represents proportion that overlap LTRs (orange), TIRs (purple), or Helitrons (green). B) Distal ACRs are categorized by chromatin pattern as K27me3, Kac, Transcribed, or Unmodified. The proportion of all distal ACRs (grey) and distal ACRs that overlap a TE (black) for each category. C) Proportion of elements containing a distal ACR (>2kb from nearest gene) classified as expressed (evidence for expression across any of the 70 tissues) or silent based on RNA-seq data from Walley et al. Elements were classified by the category of ACR present and the N for each category is shown above each bar.

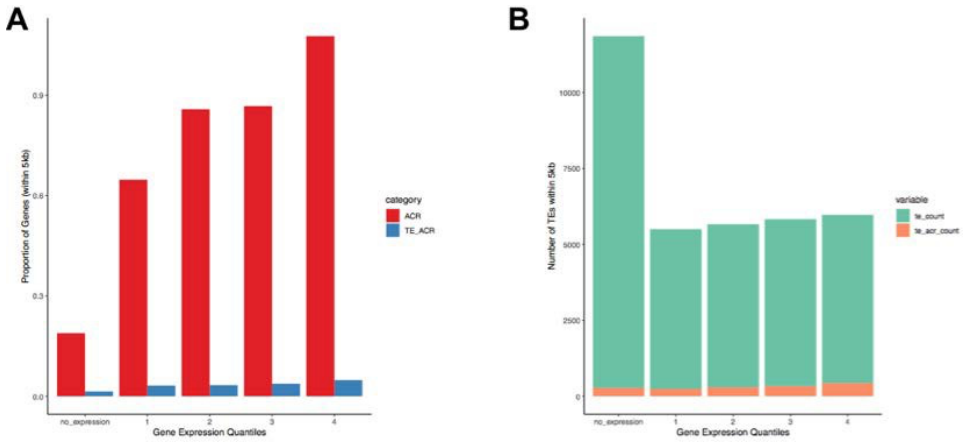

Figure S4: TE-ACRs and highly expressed genes. RNA-seq data from B73 seedling leaf samples were used to classify genes into no expression and 4 additional expression quantiles with 1 being the lowest expressed genes and 4 being the highest expressed genes (13,956 genes with no expression and 6,262 genes in each supplemental quantile). For each gene category we assessed the proportion of genes that have a defined ACR (red) or TE-ACR (blue) within 5kb (A) and the number of TEs (green) and TEs containing an ACR (orange) within 5kb of the gene (B).

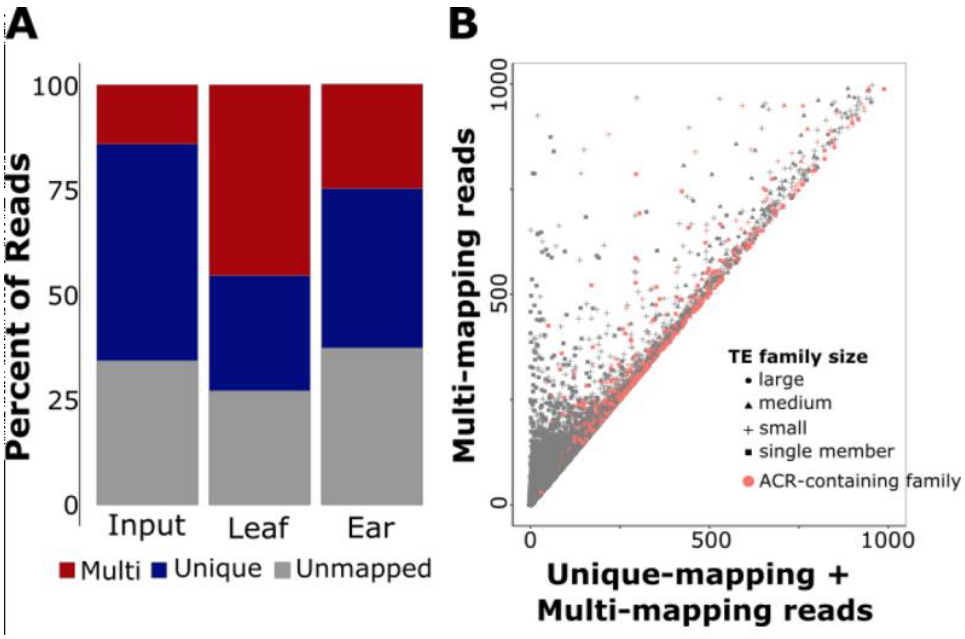

Figure S5: ATAC-seq unique and multi-mapping. A) Proportion of reads uniquely mapped, multi-mapped, or unmapped to the B73v4 genome for an input WGS dataset, ATAC-seq leaf dataset, and ATAC-seq ear dataset. B) Per family unique vs. multi- mapped read counts. Families defined by an ACR (based on unique mapping peak calling) are indicated in red.

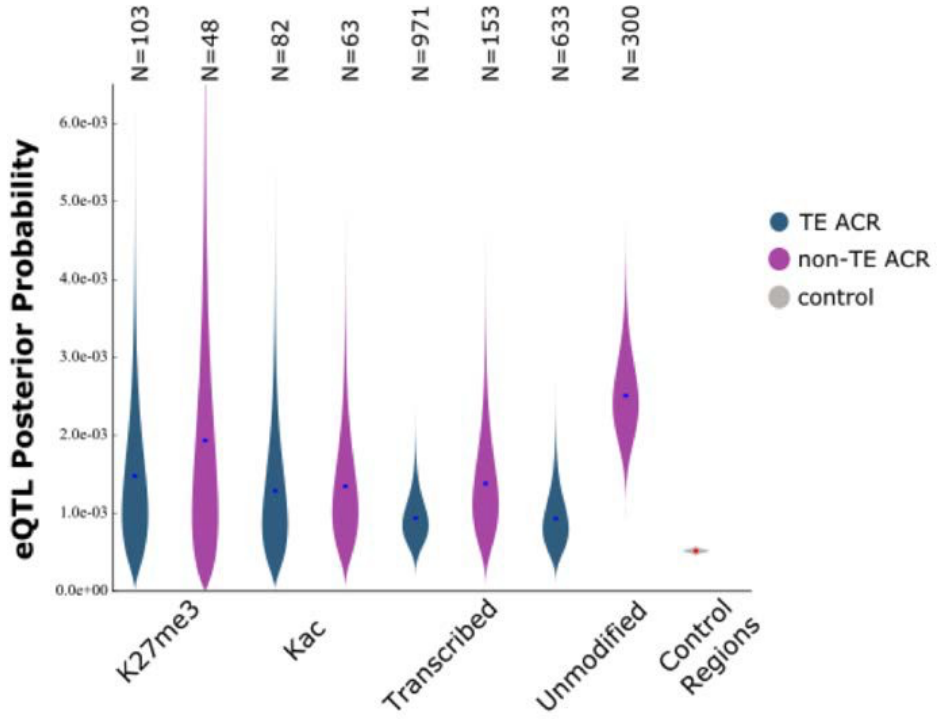

Figure S6: eQTL association. Posterior probability of association for eQTL with ACRs by chromatin class. Comparison of TE-ACRs (blue) and nonTE-ACRs (purple) to randomized control regions (grey).

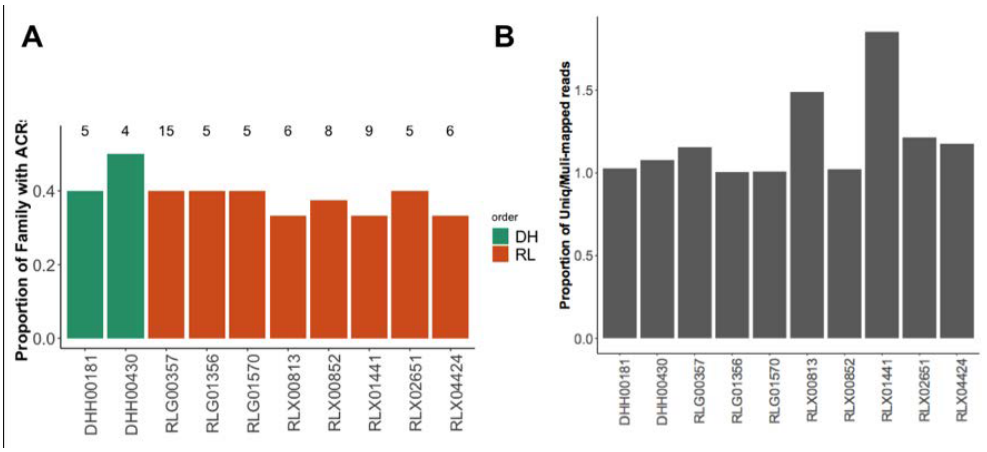

Figure S7: TE-family enrichment for ACRs. A) Subset of TE families with at least 3 members that have > 30% of their members with an ACR (based on uniquely mapped reads and peak calling). Number above bars indicates TE family size. B) Element age (by percent identity of LTR) for the LTR families with at least 3 members that have > 30% of their members with an ACR

Figure S8: Sequence similarity across members of the RLX00852 TE family. VISTA display of sequence similarity for TE family with 3 members containing an ACR (RLX00852Zm00001d00002, RLX00852Zm00001d00003, RLX00852Zm00001d00004) and 2 members lacking an ACR (RLX00852Zm00001d00001 and RLX00852Zm00001d00005). Shown relative to sequence of top TE listed. Grey boxes represent location of ACR in reference sequence.

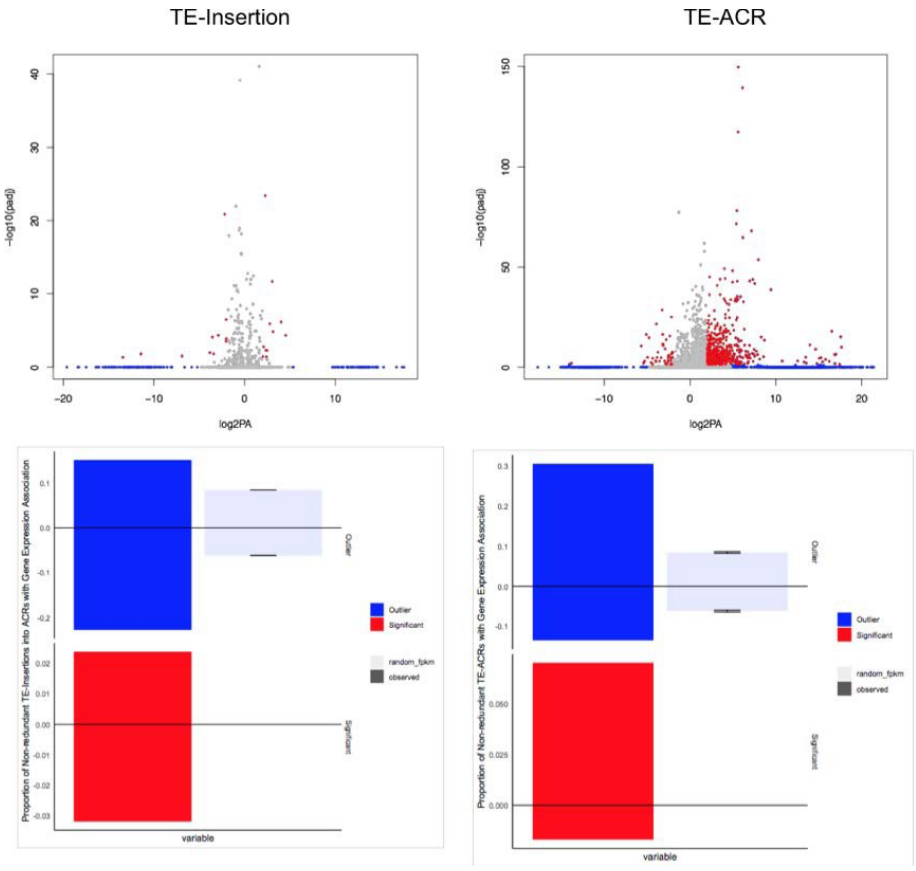

Figure S9: Combined dataset TE-Gene expression association. A/B) Volcano plot of gene expression for genes nearby B73-based ACRs with TE insertions in other genotype. (A) or B73-based TEs containing an ACR (B). A dot is present for each TE-Gene pair for RNA-seq data in each of the 8 tissues. Significant (log2(present/absent) > 2 and q- value < 0.05) and outlier (log2(present/absent) > 5) shown with red and blue points respectively. C/D) Proportion of non-redundant significant (red) or outlier (blue) expression patterns associated with TE-Insertions disrupting an ACR (C) or TE-ACRs (D).

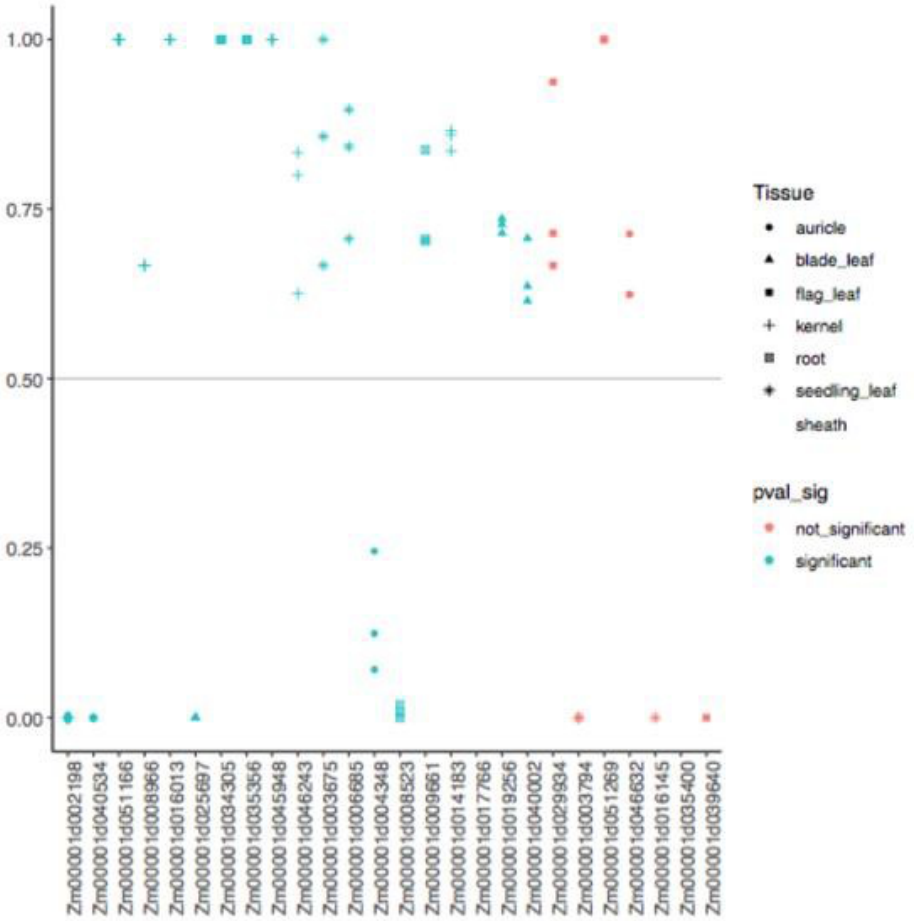

Figure S10: Allele specific expression of significant TE-Gene pairs. A subset of TE- ACRs with significant gene expression changes when present in B73 and absent in Mo17 and contain SNPs within the coding region (N=26) were assessed. Allele specific expression data from Zhou et al., (2019) was used to plot the ratio of the B73 to Mo17 alleles for the tissue determined as significant in the TE-ACR analysis. Tissue type is shown by shape and significant genes (based on a chi-square p-value < 0.1) is shown in blue.

CHAPTER V: Context Statement

Genetic variation undoubtedly influences natural epigenomic variation yet is challenging to assess without high quality reference genomes for the individuals being studied. Many studies rely upon alignments of chromatin data to a single reference genome. This limits the ability to assess the regions that are unique to one genome. The availability of multiple maize *de novo* genome assemblies provides the opportunity to study chromatin accessibility and DNA methylation in a pan-genome context. More unmethylated regions and chromatin accessible regions are discovered in B73, Mo17, Oh43 and W22 by using each genome for alignments rather than relying on a single reference genome. Chromosomal alignments reveal that we can assess only 55% of the genome that is classified as shared between any two genotypes. The unmethylated regions and accessible regions are substantially enriched within the shared portions of the genome, suggesting that the bulk of the non-shared genomic space is highly methylated and inaccessible. Shared sequence allows for the comparison of methylation level and accessibility between genotypes. The majority of unmethylated regions or accessible regions are consistent between genotypes and the subset of unmethylated regions that are present within one genotype but methylated in another are depleted for overlaps with accessible chromatin. Regions with methylation only present in one genotype were found to include various context-specific types of methylation that are associated with genomic location and have varying functional impacts. The ability to compare chromatin properties among individuals with major genomic content variation will enable pan-epigenome analyses to study the role of structural variation in influencing chromatin.

During the course of this work many authors contributed to data collection, processing, analysis, and writing. Jaclyn M Noshay, Zhikai Liang, Alexandre P Marand, Peng Zhou, and Peter A Crisp performed research. In particular, I helped process samples and data analysis. Zhikai Liang provided computational tools for accessible chromatin analysis and RNA-seq analysis. Peter A Crisp provided computational tools for characterization of methylation domain and UMR calling as well as sRNA data processing. Peng Zhou provided computational support for identification of shared and nonshared sequence for cross-genotype analysis. Figures were contributed by Jaclyn M Noshay. Texts were contributed by Jaclyn M Noshay and Nathan M Springer with discussion and editing by Robert J Schmitz and Candice N Hirsch.

CHAPTER V:

A pan-genomic analysis of maize methylomes

Introduction:

The 2.1Gb maize B73 genome was first assembled in 2009 and contains ~80% repetitive sequence which leads to regions of extensive heterochromatin (Schnable et al. 2009). Unlike model species such as *Arabidopsis thaliana*, maize has transposable elements and highly methylated regions that are interspersed with genic regions of the genome (Springer and Schmitz 2017; Baucom et al. 2009; The Arabidopsis Genome Initiative 2000). One challenge in complex crop genomes such as maize is the identification of functional elements within genomes. There are opportunities to utilize both chromatin properties such as DNA methylation and chromatin accessibility to identify functional elements.

The maize genome is highly methylated and regions containing DNA methylation can be sub-classified based on the specific sequence context of the methylation. High levels of CG and CHG methylation without CHH methylation are often found over transposable elements and other repetitive regions of the genome while CG-only methylation is observed frequently within gene bodies (Niederhuth et al. 2016; West et al. 2014; Crisp et al. 2020). CHH methylation, which is largely the result of RNA-directed DNA methylation (RdDM), is found near highly expressed genes (Gent et al. 2013; Li et al. 2015; Niederhuth et al. 2016). A small proportion of the maize genome lacks DNA methylation and these unmethylated regions (UMRs) likely reflect regions with potential roles in regulation of gene expression (Crisp et al. 2020; Ricci et al. 2019; Oka et al. 2019).

Chromatin accessibility is another feature of chromatin that can be used to identify genomic regions with roles in regulation of transcription. In maize ~1% of the genome contains open chromatin but these regions are enriched for functional significance (Rodgers-Melnick et al. 2016). Profiles of chromatin

accessibility combined with other chromatin modifications have identified potential regulatory elements in the maize genome (Oka et al. 2017; Ricci et al. 2019). While chromatin accessibility is quite useful for identifying regulatory elements in a particular tissue, this property is highly dynamic with changes between tissue types or cells (Crisp et al. 2020; Ricci et al. 2019; Marand et al. 2020). The vast majority of accessible regions occur within regions of the genome that are unmethylated. However, there are additional unmethylated regions that do not exhibit accessibility. These likely reflect the fact that the unmethylated regions of the genome are quite stable in vegetative tissues while chromatin accessibility is highly tissue-specific. To date, the analysis of chromatin accessibility in maize has largely focused on the accessible regions within the B73 genome.

The analysis of chromatin properties within the B73 reference genome has been useful for functional annotation of the genome. However, there is also value in assessing natural variation for the chromatin properties in different inbred lines of maize. While chromatin accessibility studies have largely focused on B73, many studies have compared DNA methylation between maize genotypes (Li et al. 2015; Eichten et al. 2013; Regulski et al. 2013; Anderson et al. 2018; Xu et al. 2020; Xu et al. 2019). These studies have found many examples of DNA methylation variation. Changes in DNA methylation can occur due to alterations in genetic sequence such as transposon insertions (Noshay et al. 2019) or can occur in regions with no genetic changes (Eichten et al. 2011). The ability to fully compare DNA methylation patterns among genotypes and to investigate the role of structural variation has been limited due to reliance upon a single reference genome for comparisons.

The genome content varies substantially among maize genotypes (Fu and Dooner 2002; Springer et al. 2009; Swanson-Wagner et al. 2010; Anderson et al. 2019). The availability of multiple de novo assembled reference genomes has enabled whole genome comparisons of genome content (Sun et al. 2018;

Hirsch et al. 2016; Springer et al. 2018; Haberer et al. 2020). Many of the sequences present in any one inbred are not present at collinear regions in other genomes (Sun et al. 2018; Haberer et al. 2020; Fu and Dooner 2002). This results in a pan-genome that contains more genes and transposons than any individual maize inbred (Hirsch et al. 2014; Anderson et al. 2019). While it is quite clear that genome content differs substantially it has been difficult to assess the chromatin of the pan-genome due to technical difficulties in connecting the same sequence regions between genotypes.

In this study we generated DNA methylation and chromatin accessibility profiles for four maize inbred lines that each have de novo genome assemblies. We identified the unmethylated regions (UMRs) and accessible chromatin regions (ACRs) in each of the genotypes. Chromosomal alignments were used to document shared and non-shared sequences between genomes. An assessment of the distribution of UMRs and ACRs revealed that these regions are depleted within the non-shared portions of the genome. Even within the shared regions of the genome there are examples of changes in DNA methylation or chromatin accessibility. The type of DNA methylation change is associated with the genomic properties of a locus. By assessing the stability of unmethylated regions in a pan-genomic context we find evidence that structural variation plays a role in creating changes in DNA methylation among maize lines.

Results:

Characterization of unmethylated DNA and accessible chromatin in four maize genomes

Whole genome profiles for DNA methylation (WGBS) and chromatin accessibility (ATAC-seq) were generated for seedling leaf tissue of four main inbred lines (B73, Mo17, W22 and Oh43). Two independent biological replicates of tissue were collected for each of the four genotypes and this tissue sample was used to create both WGBS and ATAC- seq libraries as well

as RNA-seq data. The resulting datasets were aligned to their own genome assembly and non-B73 genotypes were additionally aligned to the B73v4 reference.

Over 290 million reads were generated for each of the biological replicates of WGBS data and the conversion rates were >99% (Table S1). The alignment rates were substantially higher when data was mapped to the proper reference genome (~60%) as compared to when non-B73 samples were mapped to the B73 genome (~43%). The reduced mapping rate when aligning data from non-B73 genotypes to the B73 genome is likely due to polymorphisms and structural variants present between inbred lines. The data for two biological replicates were merged to calculate context-specific methylation levels for each 100bp bin resulting in ~11-13X coverage. The overall methylation state for each bin was classified based on which contexts of DNA methylation were present in that bin (Figure S1A) as described previously (Crisp et al. 2020). Bins were classified as unmethylated (<20% methylation in all contexts), CHH (CHH>15%), CG/CHG (>40% both CG and CHG), CG only (>40% CG), missing data, missing sites or intermediate methylation (Figure S1A). The majority (71-74%) of the maize genome is classified as methylated with most of this exhibiting CG/CHG methylation and quite rare CHH methylation (Figure S1A). A much smaller proportion (6-7%) of the genome is classified as unmethylated (Figure S1A). In each genome roughly 15% of the bins are classified as missing data, likely due to an inability to map WGBS reads uniquely to repetitive regions. However, if the non-B73 WGBS data is aligned to the B73 genome the proportion of bins with missing data becomes substantially larger (Figure S1B).

The unmethylated 100bp bins were merged and filtered (Crisp et al. 2020) to identify unmethylated regions (UMRs) (Table 1). UMRs were defined for each inbred based on alignment to their respective genome assembly and non-B73 samples were aligned to B73 (Figure 1A). The total number of UMRs was

similar across all 4 genotypes, although a greater number of UMRs were defined when the data was aligned to the genome from which the sample originated. There is a consistent distribution of UMRs in genic, proximal (<2kb from nearest gene) and intergenic UMRs in all four genotypes (Figure 1B). Prior studies have found that unmethylated portions of the genome often contain regulatory regions (Oka et al. 2017; Ricci et al. 2019; Crisp et al. 2020).

ATAC-seq was used to identify accessible chromatin regions (ACRs) present in these samples. ACRs were identified in each individual sample as well as using the merged biological replicates (Table S2). We focused on analysis of the ACRs called from the merged replicates since the data from the two biological replicates was highly correlated and the ACRs identified within individual samples are frequently found in the merged sample (Table S3, Figure S2). There are 21,232-24,309 ACRs present in each of the four genotypes (Table 1, Figure 1C). Relative to UMRs the ACRs are more enriched in proximal regions of the genome and depleted within intergenic regions but there is still >24% of the ACRs that are found >2kb from the nearest gene (Figure 1B). The vast majority of ACRs are found within UMRs in each of the four genotypes (Figure 1D, S3). Only a subset of the ACRs that do not overlap a UMR (86%) represent instances of methylated regions as many of these include missing methylation or low levels of methylation (Figure 1D, S3A-C). While the vast majority of ACRs occur within UMRs there are many UMRs without accessibility (Figure 1D). The UMRs could be split into accessible UMRs (aUMRs) or inaccessible UMRs (iUMRs) based on whether they overlap an ACR. The presence of an aUMR, which includes the presence of an accessible region, is much more common within or near genes that are highly expressed but is quite rare for lowly expressed genes (Figure S3D). In contrast, iUMRs are present near genes with low and high expression levels but are depleted near silent genes (Figure S3E).

Classification of shared and non-shared genomic regions

Previous studies have compared natural variation in DNA methylation based on alignment to a single reference genome (Li et al. 2015; Regulski et al. 2013). However, when data from non-B73 genotypes is mapped to the B73 genome the proportion of regions with missing data increases substantially (Figure S1B) and we do not assess the methylation levels for any genomic regions that are missing in B73. The availability of multiple reference genomes provides the opportunity to assess DNA methylation levels in the pan-genome that includes both shared (syntenic) regions of the genome as well as non-shared regions that are present in one line but missing in the other genotype. The alignment of WGBS or ATAC-seq data to the proper genome provides the advantage of more complete characterization of DNA methylation or chromatin accessibility but introduces complications for the direct comparison of specific regions between genomes.

Chromosomal alignments were performed between the B73 genome and the other reference genomes to identify the shared and non-shared genomic segments between any two genotypes (see methods for details) (Figure 2A). The approach that was implemented focused on strict criteria for identification of shared regions and the non-shared regions include both structural variants as well as highly repetitive regions that could not be uniquely mapped. Each of the other three genomes shares roughly 55% of the genome in syntenic positions relative to B73 with the remaining 45% of the genome not aligning to the B73 genome (Figure 2B). As a quality control measure we assess the proportion of space classified as shared or non-shared within identity-by-state (IBS) regions between genomes. The majority (94%) of these IBS regions are classified as shared between any two genomes (Table S4) and the regions that are not classified as shared within IBS regions are highly enriched for repetitive sequences.

Our analysis of DNA methylation or chromatin accessibility is often focused on 100bp bins. In order to directly compare the same coordinate space between genomes we identified coordinate space based on the 100bp bins from the B73 genome that are shared in each of the other genotypes (Figure 2A, S4). In the comparisons of B73 to the other three genomes we find 41-48% of the B73 bins are non-shared, 37-42% of bins have an exact match in shared regions, 12-14% mapped with >= 1 SNP, and an additional 4% mapped with >= 1 small (<20bp) indel between the two genotypes. Across all comparisons there are over 0.8 million 100bp bins that are shared in all four genotypes (Figure 2C). There are 0.5 million bins that are found only in B73 and another ~0.8 million that are present in B73 and only one or two of the other two genotypes (Figure 2C). The regions that are shared between genotypes have fewer bins with missing data such that only 6.7% of the bins shared in all three genotypes lack DNA methylation data compared to 28.4% of the bins that are only present in B73. This likely reflects the fact that much of the non-shared sequence between genomes is highly repetitive and recalcitrant to unique mapping. The identification of these shared bins allowed us to calculate the methylation levels or ATAC-seq read depth for the specific coordinates in a second genome that correspond to the B73 bins to allow direct comparisons of chromatin properties between genomes using epigenomic data aligned to its own reference genome.

UMRs are depleted in non-shared portions of the genome

We initially focused on the chromatin properties of the non-shared portions of the genome to assess the frequency of UMRs or ACRs within the pan-genome compared to the shared genome. The shared regions between genomes are generally enriched for genes while non-shared regions often have higher proportions of intergenic and TE sequence (Figure 3A, S4). The analysis of the *bz1* locus on chromosome 9 illustrates these trends of shared space in genic regions and large non-shared blocks between genes, as previously described (Wang and Dooner 2006; Fu and Dooner 2002) (Figure 3A). In the *bz1* region,

very few UMRs are found within the non-shared regions (Figure 3B) and this is also observed in a larger 58kb block on chromosome 9 in B73 from the *znf2* to the *stk1* gene annotations (Figure S4). We proceeded to perform a genome-wide assessment of the proportion of UMRs within shared and non-shared regions of the genome. Of the 107,178 UMRs identified in B73, 95% were found in shared sequence space in at least one other genotype and 75% were found in shared sequence across all three genotypes providing evidence for depletion of UMRs within non-shared sequence (Figure 3C). The unmethylated regions in B73 represent only 6% of the entire genome but are enriched within sequence which is shared. On average >12% of shared space contains UMRs while only ~2% of non-shared space contains UMRs (Figure S5). A similar analysis of the genome-wide distribution of ACRs reveals that accessible chromatin is even more enriched within genomic regions that are shared among all four genotypes (Figure 3C). ACRs account for 1.2% of the shared genomic space but only 0.1% of the non-shared genomic regions (Figure S5). The assessment of UMRs or ACRs that are found within W22, Oh43 or Mo17 genomes reveals similar trends with relatively few UMRs/ACRs in the non-shared portions of these genomes. The subset of UMRs and ACRs that are within non-shared genomic regions are features that are unique to a genome and can't be compared across genotypes. The UMRs that are present within non-shared regions are depleted for genic sequences (Figure S5B), as expected due to the depletion of genes within the non-shared regions. However, there are still a substantial number of UMRs within non-shared regions that are within genes or proximal to genes (Figure S5B). These analyses suggest that pan-genome assessment of UMRs and ACRs will provide limited discovery of novel UMRs or ACRs in non-shared space. The subsequent analysis will focus on the UMRs and ACRs that are present within shared regions of any two maize genomes.

Analysis of UMRs in shared space reveals examples of conserved and variable mC states

We proceeded to focus on the UMRs and ACRs that are present within shared regions between maize genomes. Using the coordinates of the B73-based UMRs allowed us to assess the DNA methylation state for the corresponding region in the other genomes. There are examples of UMRs that remain consistently unmethylated in the other genotype as well as examples that are methylated (Figure 4A). To assess the frequency of methylation patterns in B73-defined UMRs, we classified each B73 UMR in comparison to the methylation level observed in each of the other three genotypes (Figure 4B). The UMRs located within shared genomic regions were classified as having a consistent unmethylated state, having variable methylation such that the other genotype was classified as methylated or only having DNA methylation data for B73 (lack of WGBS coverage in the other genotype) (Figure 4B-C). The majority of B73 UMRs with data in the other genotypes were classified as having consistent UMRs but there are 6-7% of the B73 UMRs that are methylated in the other genotype (Figure 4C). The rate is much lower if we look at B73 UMRs that are present within large (>1Mb) IBS blocks (Table S4). Within these regions only 2.6% of the regions that are unmethylated in B73 are classified as methylated in another genotype. This suggests that methylation patterns are more stable in large regions that lack structural variants but may be more common in highly polymorphic regions of the genome. This analysis that begins with B73 UMRs and their coordinates can identify B73 UMRs that are methylated in other genotypes. It is likely that there is a similar number of UMRs in W22, Mo17 or Oh43 that are methylated in B73, but this would require using different coordinate space for each comparison.

Unique properties of regions with different classes of methylation change

B73 UMRs that are methylated in another genotype can be subdivided based on the prominent class of methylation in the other genotype (Figure 4D). Each of these classes of methylation likely reflect distinct mechanisms and

chromatin types. The types of methylation observed in these regions do not reflect the genome-wide proportions of methylation types (Figure S1). The proportions that are classified as CG only or CHH are higher than observed genome wide (Figure S1, 4D). While CG/CHG regions are not as common as they are genome-wide there are still many examples of CG/CHG at these regions of variable methylation (Figure 4D). In cases where multiple genotypes exhibit methylation for a region that is a UMR in B73, we found that the vast majority of them were consistently classified as the same type of methylation change. We hypothesized that regions with differing types of variable methylation would have unique properties relative to annotation, presence of accessible regions and other factors relative to the regions that are consistently unmethylated in both genotypes and proceeded to characterize the attributes of the CG only, CG/CHG and CHH variable methylation regions.

The regions that have only CG methylation in another genotype, and are unmethylated in B73, are substantially enriched within genes (Figure 5A). This is not surprising as most examples of CG only methylation occur within genes and this is often referred to as gene body methylation (gbM) (Niederhuth et al. 2016; Bewick and Schmitz 2017). Prior studies have found many examples of variable methylation between genotypes that is solely due to changes in CG methylation and these are often present within genes (Bewick and Schmitz 2017; Li et al. 2015). These CG-only variable regions very rarely overlap accessible regions in either genotype (Figure 5B). RNAseq data from the same tissue samples was used to assess gene and TE expression levels. The genes that include a B73 UMR that has CG only methylation in another genotype are enriched for genes that are stably expressed in both genotypes (Figure S6A). This is consistent with prior observations of enrichment of gbM in genes with constitutive expression (Bewick et al. 2016). Transposable elements that overlap a B73 UMR that gain CG methylation in another genotype are enriched for expression relative to other TEs (Figure S6C).

The most common DNA methylation state in the maize genome is characterized by high levels of CG and CHG methylation with very low or no CHH methylation. While the frequency of these is reduced in regions of variable methylation relative to the genome-wide frequency, there are many examples of regions that are unmethylated in B73 that are classified as CG/CHG methylation in another genotype (Figure 4D). These regions with variable CG/CHG are frequently found in intergenic regions (Figure 5A). They very rarely contain ACRs in both genotypes but do have examples of B73-specific ACRs in the unmethylated region (Figure 5B). These variable CG/CHG regions show enrichments for 22nt siRNAs in the genotype with methylation (Figure 5C). The genes that contain B73 UMRs that have CG/CHG methylation in another genotype are enriched for examples of genes with expression only in B73 and a similar trend is observed when these regions occur in gene proximal regions (Figure S6A-B).

CHH methylation, which often reflects RdDM activities, represents only 1.2% of the methylated maize genome. This type of methylation is more commonly observed in the B73 UMRs that have a variable methylation state in another genotype, representing ~10- 12% of these regions (Figure 4D). These regions that contain CHH methylation often contain high levels of CG and CHG methylation as well. The CHH variable regions are enriched within intergenic regions (Figure 5A). Very few of the B73 UMR regions that have CHH methylation in the other genotype are classified as accessible in both genotypes. However, there are a substantial number of these that have a polymorphic ACR that is only detected in B73. RdDM is typically associated with the presence of 24nt siRNAs. We compared the abundance of 21nt, 22nt and 24nt siRNAs among the genotypes for these regions with variable CHH levels (Figure 5C). Many of the regions with CHH in another genotype have higher levels of expression of 24nt siRNAs in that genotype compared to B73 (Figure 5C). Interestingly a similar trend is also observed for 22nt siRNAs but not for 21nt siRNAs (Figure 5C). The CHH changes that are located within

genes include many examples of genes that lack expression in either genotype or are only expressed in B73 (Figure S6A). When the CHH changes are in regions proximal to genes they often include examples of genes that are expressed in both genotypes and are not enriched for examples of B73-only expression (Figure S6B). These observations suggest CHH methylation near genes has little effect on expression while presence of CHH within the gene may be associated with reduced expression.

Discussion:

Zea mays, unlike many other model organisms, has a large genome containing 80% repetitive sequence and high levels of DNA methylation interspersed with functional genic and regulatory regions (Schnable et al. 2009; Jiao et al. 2017). Examination of genome structure across inbred lines have identified extensive polymorphism in both genic and repeat regions of the maize genome (Anderson et al. 2019; Springer et al. 2016; Hirsch et al. 2014; Darracq et al. 2018; Chia et al. 2012). Prior analyses of natural variation of chromatin in maize have been based on epigenome profiling data aligned to a single reference genome (Li et al. 2015; Xu et al. 2020). While a single reference genome provides insight into variation in conserved genomic regions, it does not contain the full set of sequences present in the lines being compared, resulting in biases in the ability to compare chromatin properties. The availability of multiple de novo genome assemblies allows for a more complete discovery of regions with specific chromatin properties, such as unmethylated regions (UMRs) or accessible chromatin regions (ACRs). In this study, we profiled genome-wide DNA methylation, based on alignments of data to the corresponding genome assembly, to identify the ~6% of each genome that exhibits an unmethylated state and the ~1% that is accessible. A pan-genomic analysis of UMRs and ACRs reveals the frequency of these features within both shared and non- shared genomic regions. Within the shared sequence regions, chromatin variation was identified to better understand the stability of the unmethylated portion of the genome in the absence of structural variation.

Pan-genome analyses reveal enrichment of unmethylated regions within shared sequence

Whole genome alignments between B73 and Mo17, W22, and Oh43 allowed for the identification of both shared and nonshared sequences. In a comparison of any two genomes, the non-shared sequence unique to each genome is primarily composed of highly repetitive sequences with extensive DNA methylation. UMRs are rare in these non-shared regions with the majority (95%) of B73 UMRs being found within the shared mappable sequence of at least one other genotype. This suggests that the nonshared regions are depleted for functional elements and contain relatively few UMRs and ACRs. Unmethylated regions can be further classified based on whether they overlap accessible regions. While UMRs are depleted in B73-specific sequence, this trend is even more pronounced when looking at UMRs that contain accessible chromatin. This highlights the observation that chromatin properties associated with functional sequence are depleted in the non-shared portions of the genome and assessment of the pan-genome content of UMRs or ACRs will provide limited discovery of additional regions as more genomes are evaluated.

Unmethylated regions in shared sequence can be used to assess the consistency of the unmethylated state in other genotypes. The identical-by-state regions between genomes provide an opportunity to document the variation in methylation state in large regions that are mostly devoid of sequence variation. The majority of UMRs present within IBS regions were found to maintain a consistent unmethylated state with only 2.6% exhibiting hypermethylation in another genotype. In contrast, we observed 6-7% of all UMRs in regions defined as shared sequence space exhibit hypermethylation methylation in another genotype. This suggests that the frequency of dynamic methylation states is much more common in regions near structural variants than in highly conserved regions as expected based on previous observations that transposon polymorphisms can trigger DNA methylation variation (Noshay et al. 2019).

Shared sequence UMRs with variable methylation state can be characterized by location and context

A subset of the shared sequence UMRs are not consistently unmethylated across genotypes and instead have high levels of methylation in at least one of the other three genotypes. The presence of methylation variation in the shared sequence regions allowed for characterization of attributes associated with chromatin state instability.

The majority of the maize genome has a methylation state defined by high levels of CG and CHG methylation with little or no CHH methylation. Smaller proportions of the genome are observed with only CG methylation, primarily in genic regions, or CHH methylation, often found in gene-proximal regions (Crisp et al. 2020). These different methylation states are associated with different annotated elements as well as distinct mechanisms of methylation maintenance. We sought to understand if B73 UMRs that are methylated in other genotypes show patterns associated with the type of methylation in the other genotype. Regions with CG only or CHH hypermethylation were enriched relative to genome-wide expectations. These methylation state changes occur in regions, and with consequences, expected for that methylation state. CG only methylation variation was predominantly found in genic regions and CHH methylation variation in proximal regions. While the presence of this methylation does define a region as a variable UMR, it does not have as extensive an impact on gene expression changes. Interestingly, CG only methylation found within TE sequence is associated with TEs that are expressed. The methylation difference most likely to be associated with changes in gene expression was CG/CHG methylation. These methylation differences were found to be highly stable across all genotypes such that a variably methylated region across genotypes was unmethylated in B73 and maintained the same methylation context in the other genotypes.

Characterization of relative dynamics of accessibility and methylation

The analysis of shared sequence UMRs that have consistent or variable methylation across genotypes can also identify examples of stable and unstable ACRs. Consistently unmethylated UMRs have accessibility in ~25% of cases. These accessible regions are often observed in both genotypes with some examples being unique to one genotype, with slight bias towards B73-only ACRs. These are examples of dynamic patterns of accessibility but a stable unmethylated state.

The comparison of accessibility in both genotypes for the variable UMRs revealed interesting patterns. The specific type of methylation variation is associated with the frequency of dynamic accessibility between genotypes. Genome-wide profiling of accessibility has pointed to enrichment for ACRs within gene-proximal regions (Ricci et al. 2019). The accessibility within B73 unmethylated regions with variable methylation in another genotype appears to follow this same trend. Variable CG methylation regions, predominantly found within genes, are greatly depleted for the presence of accessible regions in either genotype. In contrast, regions variable with either CG/CHG or CHH methylation variation show accessibility in the B73 genome at a rate similar to that in the consistently unmethylated regions. For these regions there is often a lack of accessibility in the genome containing methylation therefore suggesting the potential role of these regions in regulatory variation across genotypes.

While there are significant complications in comparing chromatin properties in a pan- genome aware fashion, there are also opportunities to better understand the epigenome and its variability. Unmethylated regions and accessible regions were primarily identified within shared sequence across several maize inbred lines, leading to the conclusion that the majority of these sequences with potential coding or regulatory function is captured through assessment of the shared genome. By comparing chromatin within shared sequence regions, we can isolate differences in chromatin from structural

variation such as presence/absence of the sequence. The chromatin properties of this shared sequence are often quite stable with only a subset exhibiting variability in a context-specific manner. Characterization of the epigenetic stability across maize genotypes enables deeper understanding of the sources and consequences of chromatin variation.

Methods:

Reference Genomes:

Whole genome assemblies for 4 maize inbred lines, B73 (Jiao et al. 2016), W22 (Springer et al. 2018), Mo17 (Sun et al. 2018), and Oh43 (maizegdb.org/genome/assembly/Zm-Oh43-REFERENCE-NAM-1.0) were used for genome-wide analyses. All analyses were performed on assemblies of chromosomes 1-10 while all unplaced scaffolds were disregarded due to the inability to compare these regions across genotypes. Filtered gene and structural TE annotations (Stitzer et al.; Anderson et al. 2019) were used.

Sample Collection:

Maize B73, Mo17, W22, and Oh43 plants were glasshouse grown under normal conditions. DNA was extracted from leaves of two-week old V2 plants using the DNeasy Plant Mini kit (Qiagen). Two biological replicates were sampled for sequencing and later combined into a single data set per genotype.

WGBS protocol:

1ug of DNA in 50ug of water was sheared using an Ultrasonicator to approximately 200- 350bp fragments. 20ul of sheared DNA was then bisulfite converted using the EX-DNA Methylation-Lightning Kit (Zymo Research) as per the manufacturer's instructions and eluted in a final volume of 15ul. Then 7.5ul of the fragmented bisulfite-converted sample was used as input for library preparation using the ACCEL-NGS Methyl-Seq DNA Library Kit (SWIFT Biosciences). Library preparation was performed as per the manufacturer's instructions. The indexing PCR was performed for 5 cycles.

Libraries were then pooled and sequenced on a NovaSeq 6000 in high output mode 125bp paired end reads over a single lane at the University of Minnesota Genomics Center. WGBS data generated in this study is deposited at NCBI SRA and available under accession.

Trim_glore (Martin 2011) was used to trim adapter sequences and read quality was assessed with the default parameters in paired-end read mode plus a hard clip of 20bp on each read due to SWIFT protocol specifications. Reads that passed quality control were aligned to their corresponding genome assemblies. Alignments were conducted using BSMAP-2.90(Xi and Li 2009), allowing only unique hits with up to 5 mismatches and a quality threshold of 20 (-v 5 -q 20). Duplicate reads were detected and removed using picard-tools-1.102 ("Picard Tools - By Broad Institute") and SAMtools (Li et al. 2009). Conversion rate was determined using the reads mapped to the unmethylated chloroplast genome. The resulting alignment file, merged for all samples with the same tissue and genotype, was then used to determine methylation level for each cytosine using BSMAP tools.

Methylation data summary:

Methylation levels were summarized using the bsmap methratio.py script to group by context (CG, CHG, CHH). The number of cytosines in every 100bp bin of the genome was determined and the proportion of cytosines defined as methylated was calculated. Coverage was calculated as CT / # of sites for each context. Methylation domain was classified for each 100bp bin based on the protocol described in Crisp et al. (2020) with criteria defined as a minimum site count of 2 and coverage of 3. Unmethylated regions (UMRs) were defined by grouping adjacent unmethylated bins.

ATAC-seq protocol and ACR classification:

Raw reads per sample were preprocessed with Trim_gloare. Trimmed reads were aligned to the Zea mays B73v4 genome and the genome assembly specific to

each sample using Bowtie v1.2.3 with the following parameters: “bowtie -X 1000 -m 1 -v 2 --best –strata”. Aligned reads were converted to bam files and sorted using SAMtools v1.9. Clonal duplicates were removed using Picard MarkDuplicates v2.23.3 (http://broadinstitute.github.io/picard/). Input data of maize B73 was retrieved from a previous publication and processed to obtain bam files with clonal duplicates removed. MACS2 was employed to call initial accessible chromatin regions (ACRs) with Input data as control (-c) and sample data as treatment (-t) using the following parameter "-g 2.1e9 -- keep-dup all --nomodel --extsize 147". The post-processing followed the same procedure as a prior publication (Ricci et al. 2019) to produce high-confident ACRs. Specifically, 1) Initial ACRs were split into 50 bp windows with 25 bp steps; 2) the Tn5 integration frequency in each window was calculated and normalized to the average frequency in the total genome; 3) windows with the normalized frequency greater than 25 were merged together allowing 150 bp gaps; 4) only merged regions greater than 50 bp were retained; 5) the mitochondrial or chloroplast genome from NCBI Organelle Genome Resources were removed using blast again sequences within merged ACR regions. The sites within ACRs that had the highest Tn5 integration frequency were defined as summits.

RNA-seq protocol:

Plants of B73, W22, Mo17 and Oh43 were grown under 16 h/8 h 30°C /20°C day/night for 13 days in the growth chamber of University of Minnesota. Each genotype contained five replicates, except B73 contained four replicates. RNA-seq data were generated in 150bp paired-end mode using NovaSeq 6000. B73, W22 and Mo17 reads were retrieved from the NCBI SRA accession PRJNA657262 (Liang et al. 2020) and Oh43 reads were deposited into NCBI SRA. All of the raw reads were preprocessed using Trim_galore and aligned against the B73 AGPv4 reference genome using HISAT2 v2.1.0 (cite). Gene annotations and disjoined TE annotations were utilized. Gene exon regions were subtracted from TE regions and then appended to original TE annotation to remove ambiguous mapping between genes and TEs. Reads per gene or TE was

determined using HTSeq-count v0.11.2 (Anders et al. 2015) and raw count data was input into DESeq2 (Love et al. 2014) to identify differentially expressed genes or TE elements.

The mean value for each feature (gene or TE) was calculated across the 4-5 replicates. Any feature with a mean value greater than 1 was considered "expressed". UMRs were associated with genes and TEs based on location relative to the feature. B73 UMRs which overlapped the annotated sequence coordinates within the genome being assessed were classified as "genic" or "TE". Those not overlapping a gene but within 2kb of the gene start or end were classified as "proximal".

sRNA protocol:

sRNA data were downloaded from the SRA accession SRA793603 and processed as described in Crisp et al (2020). Briefly, reads were trimmed using Trimmomatic (v0.32), mapped to the B73 v4 genome using Bowtie allowing zero mismatches. Reads were then filtered retaining 18-34nt reads, those mapping to structural RNAs removed, counts scaled by multimapping rate and normalized to counts per 5 million mapped reads. Counts were then split into 21nt, 22nt and 24nt groups and read abundance summarized in each 100bp fixed window of the genome based on the 5' position of the read. Small RNA for each size (21nt, 22nt, and 24nt) was filtered to include only the 100bp bins which had 10 counts per 5M in at least one of the samples assessed (B73, Mo17, and Oh43 leaf). Any 100bp bin with a value of zero was adjusted to 0.5 to allow for cross-genotype comparisons. The log2 fold change (B73 / nonB73) was calculated for each genotype to B73 for each size class.

Cross-genotype mapping:

Genome sequence from Mo17, W22 and Oh43 was first aligned to the B73 reference (Jiao et al. 2017) using minimap2 (Li 2018). The resulting alignments were merged and cleaned (removing overlapping alignment blocks and

alignment blocks containing assembly gaps) using in-house Perl scripts. BLAT Chain/Net tools were then used to create a single coverage best alignment net between the query genome (one of Mo17, W22 and Oh43) and the target genome (B73). Finally, a genome-wide synteny chain file was built for each genotype (against HM101), enabling downstream analyses such as variant detection and 100-bp tile liftover. Alignment pipeline and scripts are available on GitHub (https://github.com/baudisgroup/segment-liftover). Sequence was extracted for all 100bp bins in the B73 genome and aligned to Mo17, W22, and Oh43. Each bin was determined to be unmappable or mappable. Mappable bins were assigned coordinates in the non-B73 genome. The number of single nucleotide polymorphisms and insertion/deletions for each bin was calculated. Across all genotypes, only 4% of bins were found to have >= 1 insertion/deletion and 13% contained >=1 single nucleotide polymorphism. Bins with no more than 4 insertion/deletions of 20bp in size were kept for analyses of shared space. Each 100bp bin in B73 was designated as unmapped or provided matching sequence coordinates in each of the 3 other genotypes (Mo17, W22, Oh43).

Consistent and Variable UMRs:

B73 UMRs that were mappable to sequence in another genotype were further defined by methylation state in the corresponding genome. All 100bp bins within a defined UMR were assessed for the matching sequence coordinates in Mo17, W22, and Oh43. For each UMR, the proportion of bins classified as methylated (including CG, CG/CHG, and CHH methylation domains) was calculated. UMRs with >50% of the bins being methylated were defined as ”Variable mC” for the difference in methylation state from unmethylated in B73 to methylated in the non-B73 genotype. All other UMRs, showing an unmethylated state in both B73 and the non-B73 genotype assessed, were defined as “Consistent”.

B73 UMRs that are methylated in another genotype (Variable mC UMRs) were further classified by the type of methylation observed in the non-B73 genotype.

The variable UMRs were summarized by domain. The proportion of 100bp bins with a methylated domain, within the defined B73 UMR, for each methylation context was determined. Any UMR that had >50% of its methylated bins classified as a specific methylation context was declared to be variable in that context. Classification was determined first by CHH methylation, followed by CG/CHG methylation and lastly CG only methylation. Variable methylation type was defined individually for each genome based on the sequence coordinates of the B73 UMR.

Stable and Unstable ACRs:

Every B73 UMR was classified based on the accessibility of that shared sequence region within B73, Mo17, W22, and Oh43. All UMRs in B73 were defined as accessible (aUMR) or inaccessible (iUMR) based on its overlap with an accessible chromatin region in the B73 sample. For B73 aUMRs, the presence of an accessible region in the non-B73 genotypes was determined. The B73-based coordinates of the UMR in the corresponding genome were used to identify overlap with the ACRs defined in that genome. UMRs that overlap both an ACR in B73 and non-B73 genome were defined as stable ACRs. If the aUMR in B73 lacked accessibility in the non-B73 genome it was defined as B73-only ACR. Alternatively, if a UMR was inaccessible in B73 it could never be found accessible or show accessibility in the other genotype. If the iUMR lacked accessibility in the non-B73 genome, it was determined to have no ACR. If the sequence of the iUMR overlapped a defined ACR in the other genome, it was defined as a non-B73 ACR such that it was inaccessible in the B73 UMR but accessible in the shared sequence of Mo17, W22, or Oh43. The ACRs which were defined as either B73-only or nonB73-only were verified by assessing the 100bp cpm values within that region across the two genotypes.

Tables:

Table 1: UMR and ACR summary statistics

Sample Genotype	Reference Genotype	# of bins defined as Missing Data	# of bins defined as Methylated*	# of bins defined as Unmethylated	# of UMRs	# of ACRs
B73	B73v4	3511785	15064391	1325187	107178	24304
Mo17	Mo17	3649729	15566698	1385916	113838	24309
Oh43	Oh43	3096596	15719767	1445686	111261	22774
W22	W22	3322802	15315985	1369207	112253	21232

* Methylated is the combined value of bins defined as CG only, CG/CHG, and CHH

Table S1: Whole-genome bisulfite sequence mapping statistics

ID	Sample	Ref	Read Number	Total paired aligned	single	% Mapped	Conversion Rate
BN_4	B73 rep1	B73v4	300,486,928	178,021,507	25,335,060	0.5924434323	99.4426
BN_5	B73 rep 2	B73v4	350,481,634	208,208,122	28,551,115	0.5940628604	99.4452
MN_4	Mo17 rep 1	Mo17	292,007,848	179,344,346	24,884,501	0.6141764587	99.6838
MN_5	Mo17 rep 2	Mo17	314,362,827	190,423,990	26,428,480	0.6057458887	99.6736
ON_4	Oh43 rep 1	Oh43	307,821,313	182,376,885	24,609,037	0.5924764703	99.1924
ON_5	Oh43 rep 2	Oh43	361,663,168	207,190,640	30,351,407	0.5728828875	99.2545
WN_4	W22 rep 1	W22	297,457,705	182,386,478	25,971,081	0.6131509621	99.5322
WN_5	W22 rep 2	W22	296,944,325	174,763,852	28,906,243	0.5885408047	99.5391
MN_4	Mo17 rep 1	B73v4	292,007,848	128,166,520	31,441,058	0.4389146418	99.6838
MN_5	Mo17 rep 2	B73v4	314,362,827	135,492,111	33,028,873	0.4310055113	99.6736
ON_4	Oh43 rep 1	B73v4	307,821,313	133,522,189	32,404,591	0.4337652507	99.1924
ON_5	Oh43 rep 2	B73v4	361,663,168	151,595,793	37,947,708	0.4191629295	99.2545
WN_4	W22 rep 1	B73v4	297,457,705	128,861,298	29,791,109	0.4332088086	99.5322
WN_5	W22 rep 2	B73v4	296,944,325	123,719,969	31,243,067	0.4166436553	99.5391

Table S2: ATAC-seq mapping statistics

ID	Sample	Reference	Alignment Rate (%)	# of ACRs	% overlap with corresponding UMR file
BN1A	B73 control rep1 ATAC	B73v4	54.07486351	26547	0.9235318
BN2A	B73 control rep2 ATAC	B73v4	53.82690687	27052	0.9292474
MN1A	Mo17 control rep1 ATAC	Mo17	54.07720872	25764	0.9432542
MN2A	Mo17 control rep2 ATAC	Mo17	60.23229772	29371	0.9492356
ON1A	Oh43 control rep1 ATAC	Oh43	47.43187816	22440	0.9254011
ON2A	Oh43 control rep2 ATAC	Oh43	49.50660473	20485	0.9278496
WN1A	W22 control rep1 ATAC	W22	51.18143373	24564	0.928554
WN2A	W22 control rep2 ATAC	W22	51.5936334	21907	0.9245447
MN1A	Mo17 control rep1 ATAC	B73v4	37.45850796	25945	0.8996724
MN2A	Mo17 control rep2 ATAC	B73v4	40.03107982	28185	0.8953699
ON1A	Oh43 control rep1 ATAC	B73v4	33.85523242	26269	0.8839697
ON2A	Oh43 control rep2 ATAC	B73v4	34.5260282	24044	0.8802196
WN1A	W22 control rep1 ATAC	B73v4	34.86699401	23889	0.8814517
WN2A	W22 control rep2 ATAC	B73v4	34.66699222	22616	0.8685444

Table S3: Correlation between ATAC-seq replicates

	Correlation between replicates	% of rep1 ACRs in merged set	% of rep2 ACRs in merged set
B73	0.9652805	85%	82%
Mo17	0.9706562	92%	87%
W22	0.9662455	83%	91%
Oh43	0.9568982	91%	86%

Table S4: IBS regions between B73 and Mo17/W22

B73 coordinates	**Mb size**	**IBD genotype**	**% of bins Shared**	**# B73 UMRs**	**# Variable mC**
2:124400000-129750000	5.35	Mo17	99.55%	70	0 (0%)
2:184400000-190350000	5.95	Mo17	94.86%	359	8 (2.2%)
8:148050000-155450000	7.40	Mo17	96.17%	546	9 (1.6%)
5:118050000-123550000	5.50	Mo17	96.30%	105	7 (6.7%)
2:122450000-129750000	7.30	W22	98.67%	105	2 (1.9%)
3:67800000-85500000	17.7	W22	96.99%	264	7 (2.7%)
7:28450000-41350000	12.9	W22	96.34%	448	15 (3.3%)
1:126100000-135950000	9.85	W22	94.67%	124	3 (2.4%)
3:102650000-111700000	9.05	W22	98.11%	250	3 (1.2%)
6:44900000-51850000	6.95	W22	98.09%	129	7 (5.4%)
6:52550000-60900000	8.35	W22	96.44%	189	8 (4.2%)
8:123400000-139800000	16.4	W22	95.89%	849	21 (2.5%)

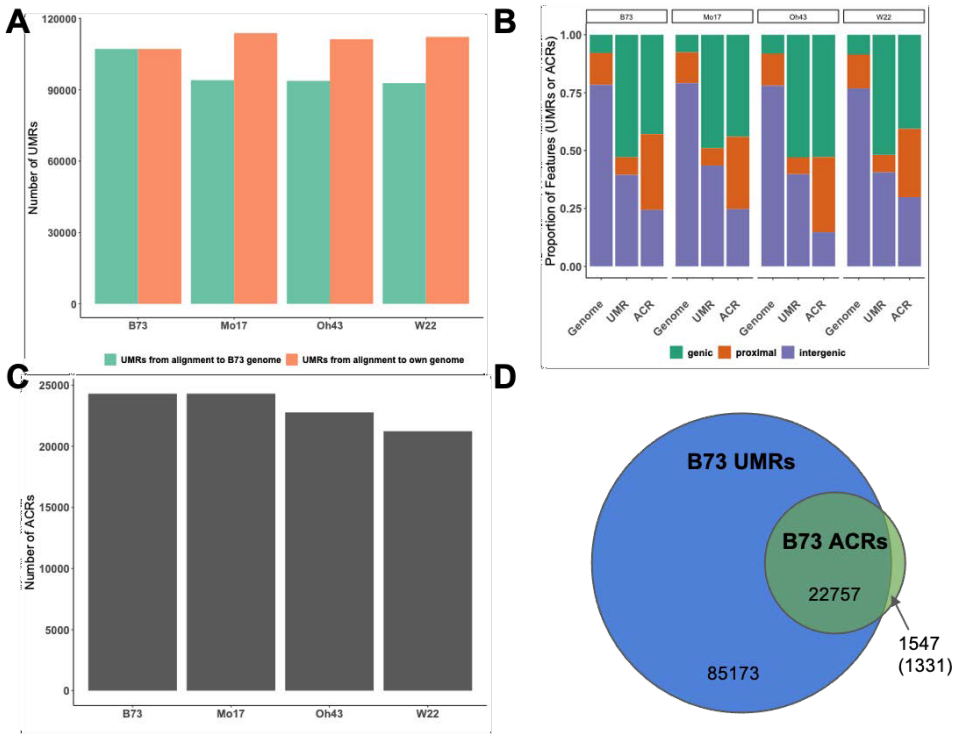

Figure 1: Summary numbers and locations. A) Number of UMRs defined based on samples aligned to B73v4 (green) and their own genome assembly (orange). B) Location of UMRs and ACRs in the genome based on gene annotations with UMRs overlapping genes (green), within 2kb of a gene (orange) and >2kb from a gene (purple). C) Number of ACRs defined based on the merged replicates for each genotype aligned to their respective genome assemblies. D) Overlap between the B73 UMRs and ACRs defined based on alignments to the B73v4 genome. Number in parentheses indicates ACRs that are defined by methylated domains.

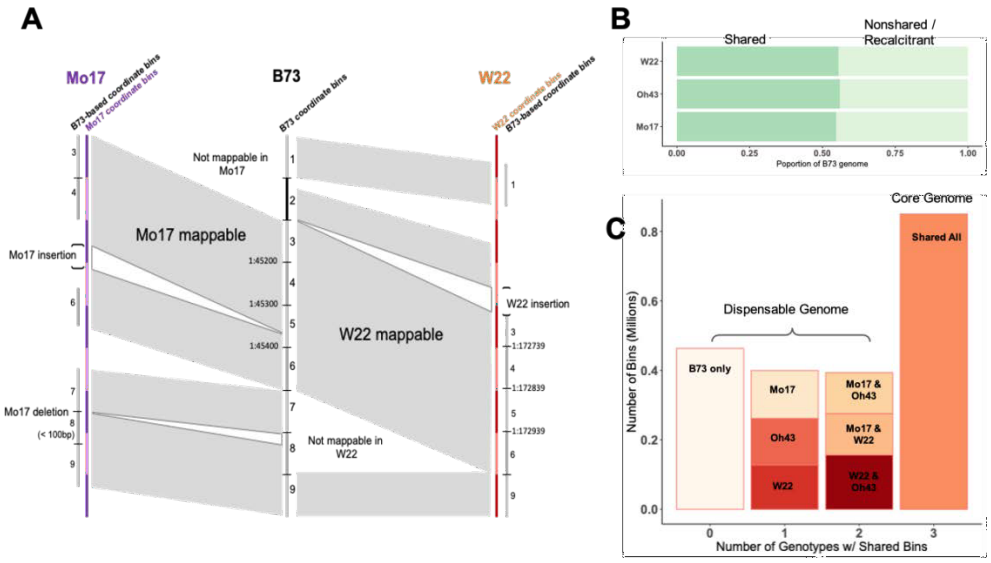

Figure 2: Defining shared and nonshared regions. A) Schematic representation of B73-based 100bp bins defined as shared or nonshared in Mo17 and W22 (gray shaded regions) based on chromosomal alignments with minimap2. 100bp bins in W22 or Mo17 could be defined by 100bp increments within that genome sequence or based on coordinate matches to the B73 genome and these are shown as the W22/Mo17 coordinate bins or the B73-based coordinates. B) Proportion of the B73 genome that is defined as shared or non-shared with Mo17, W22, and Oh43 based on chromosome-level sequence alignments. C) The proportion of B73 100bp bins that are unique to B73 (0 shared genotypes), shared with one other genotype assessed (1), shared with two other genotypes assessed (2) or shared across all 4 genotypes including B73, Mo17, Oh43, and W22 (3).

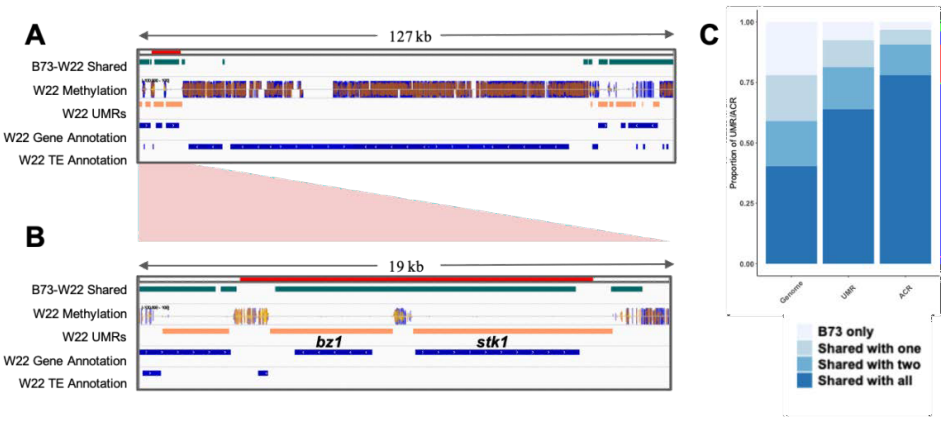

Figure 3: B73-based shared sequence bins. A) IGV representation of a 127kb segment on chromosome 9 of the W22 genome assembly showing sequence shared with B73 (green) and unique to W22, W22 methylation across all contexts (CG - blue, CHG - red, CHH - yellow), UMRs defined in W22 (yellow), and gene and TE annotations (blue). B) A zoomed in region of the bz1 locus. C) Proportion of defined B73 genome, defined UMRs and defined ACRs that share sequence with 0,1,2, or 3 of the genotypes assessed.

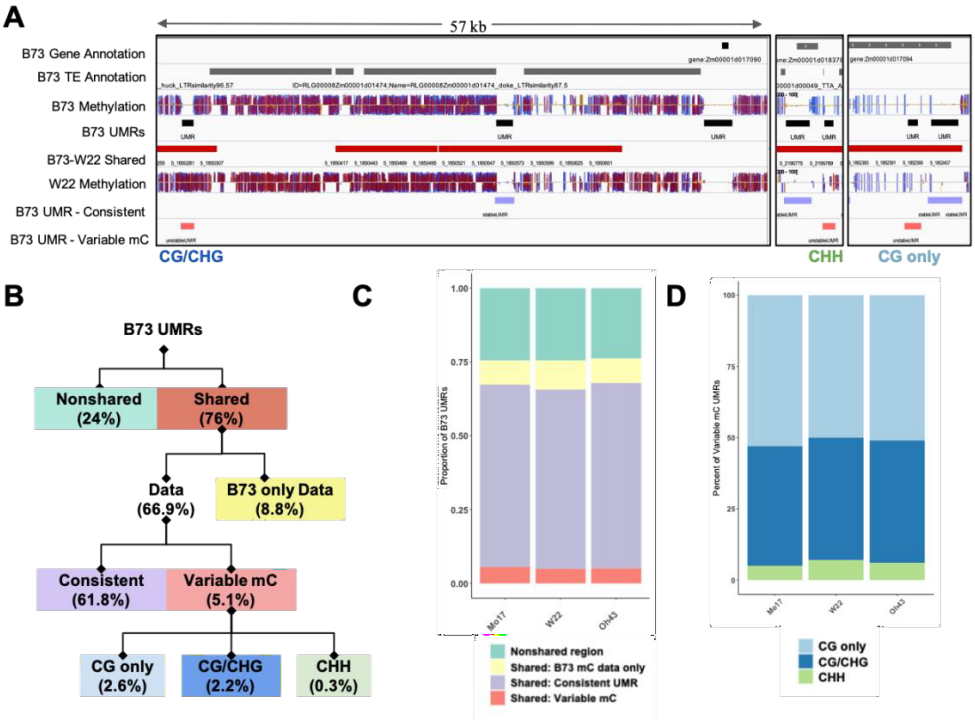

Figure 4: Stability of UMRs in shared sequence. A) Genome browser view of a region on chromosome 5 of the B73 genome. A track of B73 methylation in all contexts (CG- blue, CHG-red, CHH-yellow) is shown with UMRs defined below in black. Regions with shared sequence with W22 are shown in red and the W22 methylation track (aligned to the B73v4 assembly) with corresponding UMR classification as consistent (purple) or variable (red). Three separate snapshots are shown with the type of methylation found in W22 for the variable UMR noted below (CG only, CG/CHG, or CHH). B) Flowchart of process through B73 UMR classification with percent of all UMRs in each category listed in parentheses. C) Proportion of B73 UMRs that are shared or non-shared (green) based on sequence with the respective genome assembly. Shared regions are further classified as B73-only (yellow) for UMRs that lack data in the other genome, consistent (purple) for UMRs that maintain an unmethylated state in at least ⅔ of the corresponding bins in the other genome, or variable mC (red) for UMRs that change to a methylated state in the other genome. D) All B73 UMRs classified as variable mC (red in B) were assessed for methylation type. 100bp bin domains were examined and the UMR was classified by the majority domain. The percent of all B73 UMRs classified as variable mC that change to CG only (light blue), CG/CHG (dark blue), or CHH (green) methylation in the other genotype was calculated.

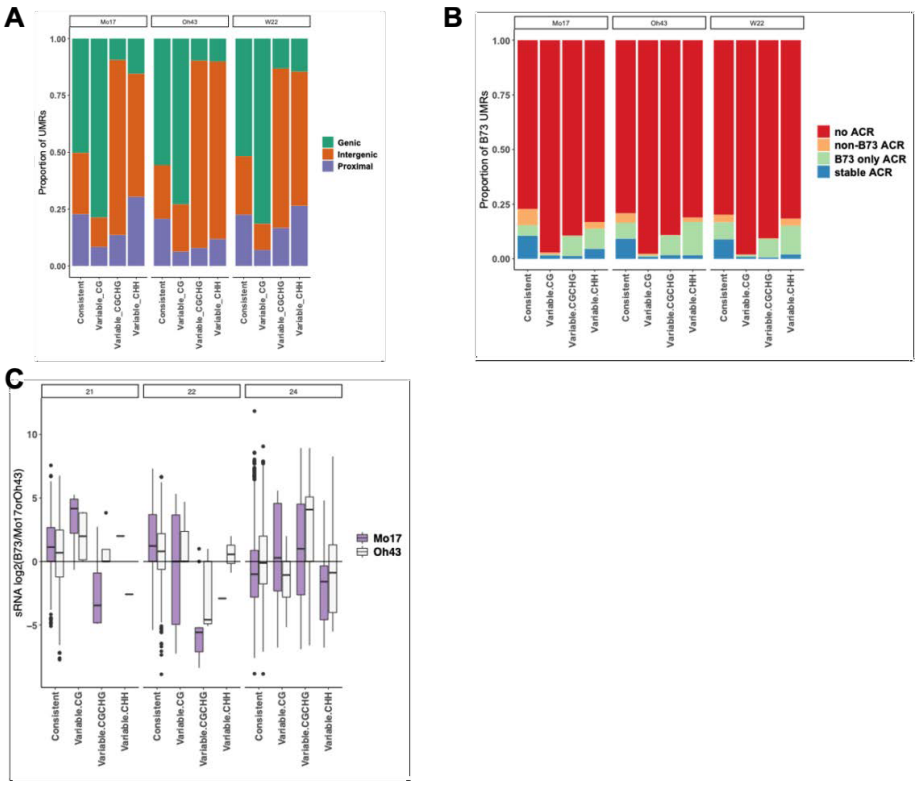

Figure 5: Characteristics of variable UMRs. B73 UMRs were broken into their methylation category (as shown in Figure 4B/C) defined as consistent, variable with CG methylation, variable with CG/CHG methylation, and variable with CHH methylation. The proportion of each UMR category was assessed for location relative to genes (A), presence of ACRs (B), and sRNA differences (C). A) UMRs were classified as genic (green), proximal (purple), or intergenic (orange) for shared space between B73 and each non-B73 genotype. B) UMRs were defined as containing an ACR in both genotypes (stable ACR: blue), in one genotype (B73 only: green, non-B73 only: orange), or lacking an ACR in both genotypes (no ACR: red). C) sRNA counts were calculated for 100bp bins of the B73 genome. Bins were filtered to contain only those with at least 10 counts per 5M in at least one of the samples (B73, Mo17, or Oh43). All sRNA values of 0 were adjusted to 0.5 to allow calculation of differences. The log2 fold change for each matching bin between B73 and Mo17 or Oh43 was calculated for 21, 22, and 24 sRNAs.

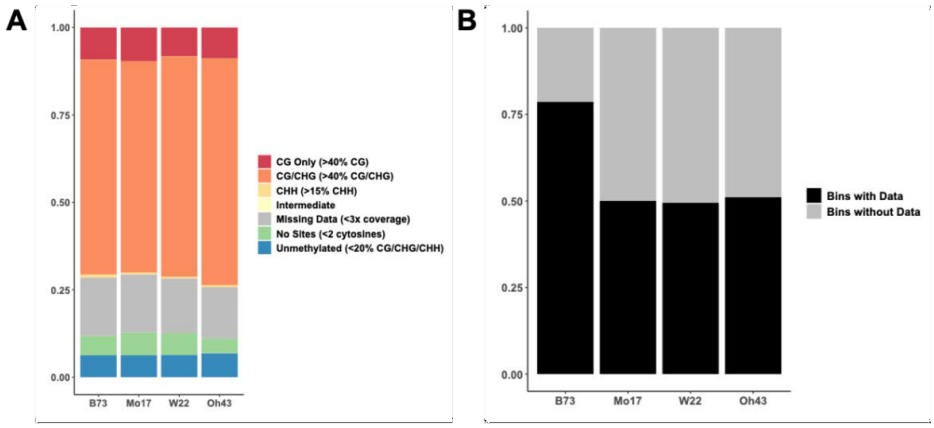

Figure S1: A) Methylation domains across samples aligned to their respective genome assemblies. Each 100bp bin of the genome was assigned a methylation domain based on CG, CHG, and CHH methylation. Any bin with less than 2 cytosines was labeled "No Sites" and any bin with < 3x coverage was labeled "Missing Data". For all other bins, context-specific cutoffs of methylation were used to classify CHH, CG only, CG/CHG, Intermediate and Unmethylated status. The proportion of each domain category for all bins in the respective genome are shown. B) The proportion of all bins aligned to the B73v4 genome, for the B73 and non-B73 (Mo17, W22, Oh43) tissue samples, thatresults in data that can be assessed (black) or bins without enough coverage to be assessed (grey).

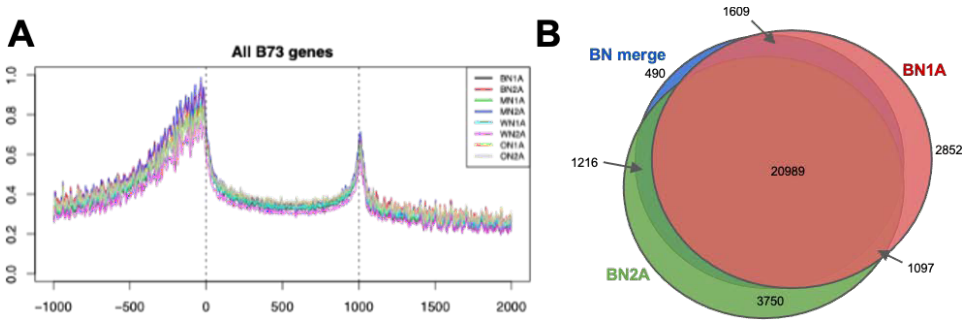

Figure S2: B73 ATAC-seq reproducibility. A) Metaplot over annotated B73 genes for all ATAC-seq tissue samples aligned to the B73v4 genome assembly. The gene space was normalized to a 1kb region (represented in the middle of the metaplot) with the flanking upstream and downstream 1kb based on gene transcript direction. B) ATAC-seq was performed on two replicates for each genotype and ACR calls were generated for each sample individually as well as the merged alignment file. The venn diagram represents the overlap in defined ACRs for individual and merged samples for B73.

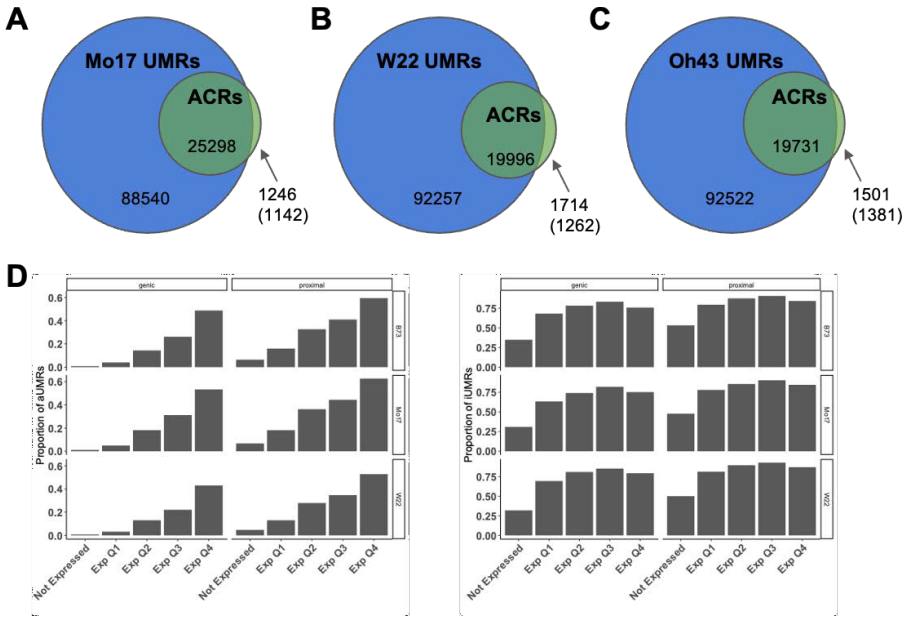

Figure S3: Overlap between the Mo17 (A), W22 (B) and Oh43 (C) UMRs (blue) and ACRs (green) defined based on alignments to the corresponding genome. Non-UMR ACRs that are defined as methylated are shown in parentheses below ACR count. D) All B73 genes were characterized by expression profile across 5 replicates as not expressed or expressed with level of expression broken into quantiles of lowest expression (Q1) to highest expression (Q4). For each category of gene, the proportion of aUMRs (left) and iUMRs (right) that are overlapping or proximal (<2kb) was calculated.

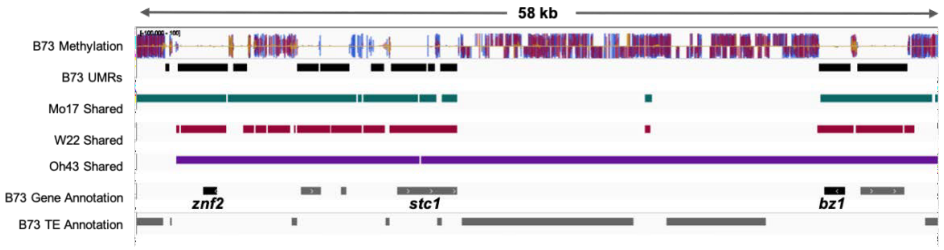

Figure S4: B73 genome sequence that is shared and nonshared in a 58kb region of chromosome 9. Tracks show B73 methylation levels in all contexts (CG-blue, CHG-red, CHH-yellow), define B73 UMRs (black), Mo17 shared sequence (green), W22 shared sequence (red), Oh43 shared sequence (purple), and B73 gene and TE annotations (grey).

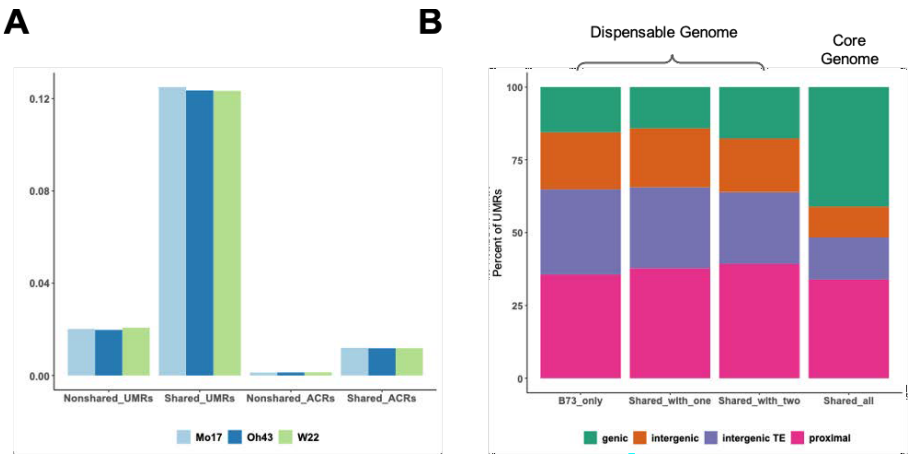

Figure S5: A) Proportion of all shared or nonshared B73 sequence that is defined as UMR or ACR in Mo17 (light blue), Oh43 (dark blue) and W22 (green). B) Percent of B73 UMRs that are defined by sequence unique to B73 (B73 only) or shared with other genotypes by location. Location was defined by UMRs that overlap a gene (green), are within 2kb of a gene (pink), are >2kb from a gene and overlap a TE (purple), or >2kb and don't overlap an annotation (orange).

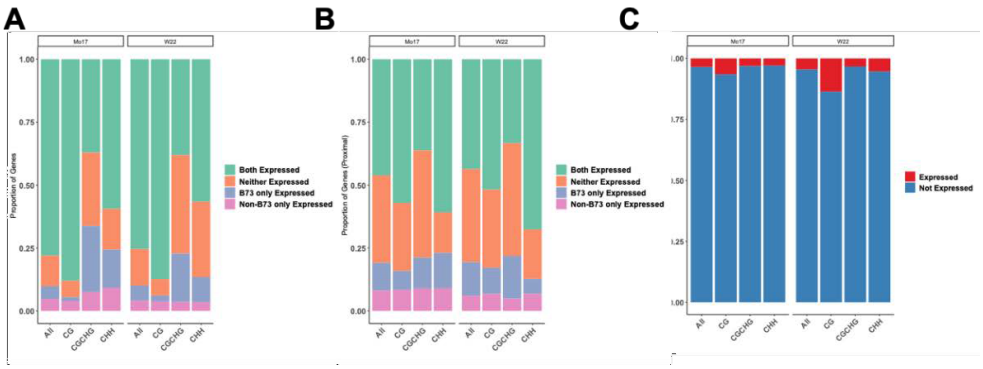

Figure S6: Features were defined as expressed (mean expression of replicates > 1) or not expressed. Genes were classified as expressed in both B73 and the non-B73 genotype, not expressed in both, expressed in B73 only, or expressed only in the non-B73 genotype. TEs were classified as expressed or not expressed in the non-B73 genotype. UMRs were broken into all defined UMRs, and variable UMRs by methylation context. A) The proportion of genes with UMRs overlapping gene annotation coordinates. B) The proportion of genes with UMRs proximal (within 2kb). C) The proportion of TEs with UMRs overlapping TE annotation coordinates.

Conclusion

Throughout my dissertation I have worked to address questions regarding the influence of repetitive sequence and DNA methylation on phenotype. I sought to understand the complex interaction between sequence variation, focusing on highly methylated transposable elements, and epigenetic architecture. The following will discuss major takeaways drawn from each research chapter of my dissertation and bring to light additional avenues for further research and application.

Chapter III addresses the interaction between transposable elements and DNA methylation in a genome containing extensive repeat sequence interspersed with coding sequence. Repetitive regions defined by annotated TEs have been classified into families based on sequence and structure. By assessing the DNA methylation trends surrounding each of these groups we were able to identify varying chromatin patterns. Cross-genotype analyses were conducted on polymorphic TE sequence to question the causative factor in each instance of TE and DNA methylation variation. TEs were found to have both chromatin preference for insertion site, with many TE-absent haplotypes showing high levels of DNA methylation, as well as influence on chromatin after insertion demonstrated by hypermethylation in the TE-present haplotype. This bi-directional interaction between chromatin and sequence shows the complex relationship between DNA methylation as a mechanism to silence a potentially disruptive sequence and the evolutionary adaptation of TEs to maintain within the genome.

Chapter IV focuses on the potential impact of transposable element presence in a genome. Two avenues of influence were assessed; the potential for TEs to create regulatory regions through introduction of novel sequence and the potential for TEs to alter accessible chromatin regions through disruption of sequence upon insertion into the genome. New resources, accessible chromatin data and sequence polymorphism calls across multiple genotypes, allowed for the discovery that TEs both disrupt and introduce putative regulatory sequences on a genome-wide scale. This is essential to understanding the role of TEs and their potential in altering the regulatory landscape of the genome. We further assessed these TE insertions with respect to their influence on nearby gene expression. It was observed that a majority of TE insertions have no significant impact on expression variation, but a subset of examples show enhancer potential within TE sequence resulting in altered gene expression in the TE present/absent haplotypes.

Chapter V looks at the subset of the genome defined by low levels of DNA methylation and high levels of accessibility. Previous chapters introduced the role of structural variation in chromatin state variation. These final studies further explored the presence of unmethylated regions and accessible chromatin in shared and nonshared genomic sequence across 4 maize inbred lines. These regions of interest were identified almost exclusively in sequence

shared between at least two genotypes and therefore suggests that regulatory and coding sequence is contained within shared sequence. Unmethylated regions in shared space allowed for the analysis of chromatin variation. The chromatin structure of these regions was maintained a majority of the time across genotypes suggesting the importance of a conserved unmethylated state. There were several examples of unmethylated regions defined in one genotype with variable methylation in another genotype. These cases were observed in all methylation contexts and were found in regions of the genome and with consequences associated with the specific context of methylation present. Initial findings present the importance of shared sequence and the ability to assess the maize pan-epigenome. Further analyses will lead to deeper understanding of sequence variation roll in epi-allele discovery.

The study of DNA methylation and transposable elements in maize has begun to address questions regarding the role of structural variation in epigenome architecture. Resources generated and discoveries presented in this dissertation allow for further examination and separation of the frequency of causation to effect in epigenetic and genetic variation across maize genotypes.

BIBLIOGRAPHY

Alleman, M. et al. An RNA-dependent RNA polymerase is required for paramutation in maize. Nature. 2006; 442(7100), pp.295–8.

Anderson S. N., M. C. Stitzer, P. Zhou, J. Ross-Ibarra, C. D. Hirsch, et al., 2019b Dynamic patterns of transcript abundance of transposable element families in maize. G3: Genes, Genomes, Genetics 9: 3673–3682.

Anderson SN, Stitzer MC, Brohammer AB, Zhou P, Noshay JM, Hirsch CD, et al. Transposable elements contribute to dynamic genome content in maize The Plant Journal. 2019; doi:10.1101/547398

Anderson SN, Zynda G, Song J, Han Z, Vaughn M, Li Q, et al. Subtle Perturbations of the Maize Methylome Reveal Genes and Transposons Silenced by Chromomethylase or RNA-Directed DNA Methylation Pathways. G3. 2018; doi:10.1534/g3.118.200284

Anders, S., P. T. Pyl, and W. Huber. HTSeq--a Python Framework to Work with High- Throughput Sequencing Data. Bioinformatics 31 (2): 166–69.

Barber, W.T. et al. Repeat associated small RNAs vary among parents and following hybridization in maize. *Proceedings of the National Academy of Sciences of the United States of America*. 2012; 109(26), pp.10444–10449.

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, et al. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genet. 2009;5: e1000732.

Bauer, M.J. & Fischer, R.L. Genome demethylation and imprinting in the endosperm. *Current Opinion in Plant Biology.* 2011; 14(2), pp.162–167.

Becker C, Hagmann J, Muller J, Koenig D, Stegle O, Borgwardt K, et al. Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature. 2011;480: 245– 249.

Bennetzen J. L., 2000 Transposable element contributions to plant gene and genome evolution. Plant Mol. Biol. 42: 251–269.

Bennetzen J. L., and E. A. Kellogg, 1997 Do Plants Have a One-Way Ticket to Genomic Obesity? Plant Cell 9: 1509–1514.

Bennetzen JL, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol. 2014;65: 505–530.

Bewick AJ, Schmitz RJ. Gene body DNA methylation in plants. Curr Opin Plant Biol. 2017;36: 103–110.

Bewick, A.J. et al. On the origin and evolutionary consequences of gene body DNA methylation. *Proceedings of the National Academy of Sciences of the United States of America*. 2016; 113(32), pp.9111–6.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30: 2114–2120.

Bond DM, Baulcombe DC. Epigenetic transitions leading to heritable, RNA-mediated de novo silencing in Arabidopsis thaliana. Proc Natl Acad Sci U S A. 2015; 112: 917– 922.

Briggs W. H., M. D. McMullen, B. S. Gaut, and J. Doebley, 2007 Linkage mapping of domestication loci in a large maize teosinte backcross resource. Genetics 177: 1915– 1928.

Brink, R.A. A Genetic Change Associated with the R Locus in Maize Which Is Directed and Potentially Reversible. *Genetics*. 1956; 41(6), pp.872–889.

Castelletti S., R. Tuberosa, M. Pindo, and S. Salvi, 2014 A MITE Transposon Insertion Is Associated with Differential Methylation at the Maize Flowering Time QTL Vgt1. G3 . https://doi.org/10.1534/g3.114.010686; 10.1534/g3.114.010686

Catoni M, Griffiths J, Becker C, Zabet NR, Bayon C, Dapp M, Lieberman-Lazarovich M, Weigel D, Paszkowski J: DNA sequence properties that predict susceptibility to epiallelic switching. *EMBO J* 2017, doi:e201695602 [pii].

Cavrak V. V., N. Lettner, S. Jamge, A. Kosarewicz, L. M. Bayer, et al., 2014 How a retrotransposon exploits the plant's heat stress response for its activation. PLoS Genet. 10: e1004115.

Chandler, V.L. Paramutation: From Maize to Mice. Cell. 2007; 128(4), pp.641–645.

Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet. 2012;44: 803–807.

Choi JY, Purugganan MD. Evolutionary Epigenomics of Retrotransposon-Mediated Methylation Spreading in Rice. Mol Biol Evol. 2018; 35: 365–382.

Chuong E. B., N. C. Elde, and C. Feschotte, 2017 Regulatory activities of transposable elements: from conflicts to benefits. Nat. Rev. Genet. 18: 71–86.

Clark R. M., T. N. Wagler, P. Quijada, and J. Doebley, 2006 A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat. Genet. 38: 594–597.

Coe, E.H. A Regular and Continuing Conversion-Type Phenomenon at the B Locus in Maize. Proceedings of the National Academy of Sciences of the United States of America. 1959; 45(6), pp.828–832.

Coe, E.H. The origins of maize genetics. Nature Reviews Genetics. 2001; 2(11), pp.898– 905.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature. 2008; 452: 215–219.

Colicchio JM, Kelly JK, Hileman LC: Parental experience modifies the Mimulus methylome. *BMC Genomics* 2018, 19:746.

Cortijo S, Wardenaar R, Colome-Tatche M, Gilly A, Etcheverry M, Labadie K, Caillieux E, Hospital F, Aury JM, Wincker P, et al.: Mapping the epigenetic basis of complex traits. *Science* 2014; 343:1145–1148.

Crisp, P.A. et al. Reconsidering plant memory: Intersections between stress recovery, RNA turnover, and epigenetics. Science Advances. 2016; 2(2), p.e1501340.

Crisp PA, Marand AP, Noshay JM, Zhou P, Lu Z, Schmitz RJ, et al. Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes. Proc Natl Acad Sci U S A. 2020. doi:10.1073/pnas.2010250117

Cuerda-Gil D, Slotkin RK. Non-canonical RNA-directed DNA methylation. Nature plants.
2016; 2: 16163.

Daccord N, Celton J-M, Linsmith G, Becker C, Choisne N, Schijlen E, van de Geest H, Bianco L, Micheletti D, Velasco R, et al.: High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. Nat Genet 2017; 49:1099–1106.

Darracq A, Vitte C, Nicolas S, Duarte J, Pichon J-P, Mary-Huard T, et al. Sequence analysis of European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. BMC Genomics.
2018;19: 119.

Denkena J, Johannes F, Colomé-Tatché M: Region-level Epimutation Rates in Arabidopsis thaliana. 2020, doi:10.1101/2020.08.18.255919.

Dietrich CR, Cui F, Packila ML, et al. Maize Mu transposons are targeted to the 5' untranslated region of the gl8 gene and sequences flanking Mu target-site duplications exhibit nonrandom nucleotide composition throughout the genome. Genetics. 2002;160(2):697-716.

Doebley J., A. Stec, and L. Hubbard, 1997 The evolution of apical dominance in maize. Nature 386: 485–488.

Du J, Zhong X, Bernatavichute YV, Stroud H, Feng S, Caro E, et al. Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. Cell. 2012;151: 167–180.

Du, J. et al. DNA methylation pathways and their crosstalk with histone methylation. *Nature*. 2015; 33(4), pp.395–401.

Du, J. et al. Mechanism of DNA methylation-directed histone methylation by KRYPTONITE. *Molecular Cell*. 2014; 55(3), pp.495–504.

Dubin MJ, Zhang P, Meng D, Remigereau MS, Osborne EJ, Casale FP, Drewe P, Kahles A, Jean G, Vilhjalmsson B, et al.: DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. *Elife* 2015, 4:e05255.

Eggleston, W.B., Alleman, M. & Kermicle, J.L. Molecular organization and germinal instability of R-stippled maize. *Genetics*. 1995; 141(1), pp.347–360.

Eichten SR, Briskine R, Song J, Li Q, Swanson-Wagner R, Hermanson PJ, et al. Epigenetic and genetic influences on DNA methylation variation in maize populations. Plant Cell. 2013;25: 2783–2797.

Eichten SR, Ellis NA, Makarevitch I, Yeh CT, Gent JI, Guo L, et al. Spreading of heterochromatin is limited to specific families of maize retrotransposons. PLoS Genet. 2012;8: e1003127.

Eichten SR, Schmitz RJ, Springer NM. Epigenetics: Beyond Chromatin Modifications and Complex Genetic Regulation. Plant Physiol. 2014;165: 933–947.

Eichten SR, Springer NM: Minimal evidence for consistent changes in maize DNA methylation patterns following environmental stress. *Front Plant Sci* 2015, 6:308.

Eichten SR, Stuart T, Srivastava A, Lister R, Borevitz JO: DNA methylation profiles of diverse Brachypodium distachyon align with underlying genetic diversity. *Genome Res* 2016, 26:1520–1531.

Eichten, S.R. et al. Heritable epigenetic variation among maize inbreds. *PLoS Genetics*. 2011; 7(11).

Feng, S., Jacobsen, S.E. & Reik, W. Epigenetic reprogramming in plant and animal development. *Science*. 2010; 330(6004), pp.622–627.

Feschotte C., 2008 Transposable elements and the evolution of regulatory networks. Nature reviews. Genetics 9: 397–405.

Forestan, C., Farinati, S., Aiese Cigliano, R., Lunardon, A., Sanseverino, W., & Varotto, S.Maize RNA PolIV affects the expression of genes with nearby TE insertions and has

a genome-wide repressive impact on transcription. BMC Plant Biology. 2017; 17, 161.

Fu H, Dooner HK. Intraspecific violation of genetic collinearity and its implications in maize. Proc Natl Acad Sci U S A. 2002;99: 9573–9578.

Furci L, Jain R, Stassen J, Berkowitz O, Whelan J, Roquis D, Baillet V, Colot V, Johannes F, Ton J: Identification and characterization of hypomethylated DNA loci controlling quantitative resistance in Arabidopsis. *Elife* 2019, 8.

Gallego-Bartolomé J, Gardiner J, Liu W, Papikian A, Ghoshal B, Kuo HY, Zhao JM-C, Segal DJ, Jacobsen SE: Targeted DNA demethylation of the Arabidopsis genome using the human TET1 catalytic domain. *Proc Natl Acad Sci U S A* 2018, 115:E2125–E2134.

Ganguly DR, Crisp PA, Eichten SR, Pogson BJ: The Arabidopsis DNA Methylome Is Stable under Transgenerational Drought Stress. *Plant Physiol* 2017, 175:1893–1912.

Gehring, M. Genomic imprinting: insights from plants. Annual Review of Genetics. 2013; 47, pp.187–208.

Gehring, M., Bubb, K.L. & Henikoff, S. Extensive Demethylation of Repetitive Elements During Seed Development Underlies Gene Imprinting. *Science*. 2009; 324(5933), pp.1447–1451.

Gehring, M., Missirian, V. & Henikoff, S. Genomic Analysis of Parent-of-Origin Allelic Expression in Arabidopsis thaliana Seeds. *PloS one*. 2011; 6(8), p.e23687.

Gent JI, Ellis NA, Guo L, Harkess AE, Yao Y, Zhang X, et al. CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. Genome Res. 2013;23: 628–637.

Ghoshal B, Vong B, Picard CL, Suhua F, Tam JM, Jacobsen SE: A viral guide RNA delivery system for CRISPR-based transcriptional activation and heritable targeted DNA demethylation in Arabidopsis thaliana. *Cold Spring Harbor Laboratory* 2020, doi:10.1101/2020.07.09.194977.

Gouil, Q. & Baulcombe, D.C. (2016) DNA Methylation Signatures of the Plant Chromomethyltransferases. *PLOS Genetics*, 12(12), p.e1006526.

Graaf A van der, Wardenaar R, Neumann DA, Taudt A, Shaw RG, Jansen RC, et al. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. Proc Natl Acad Sci U S A. 2015;112: 6676–6681.

Greenberg, M.V.C. et al. Interplay between Active Chromatin Marks and RNA-Directed DNA Methylation in Arabidopsis thaliana. *PLoS Genetics*. 2013; 9(11).

Guo, C., Spinelli, M., Ye, C., Li, Q. Q., & Liang, C. (2017). Genome-Wide Comparative Analysis of Miniature Inverted Repeat Transposable Elements in 19 Arabidopsis thaliana Ecotype Accessions. Scientific reports, 7(1), 2634. https://doi.org/10.1038/s41598-017-02855-1

Haag, J.R. et al. Functional diversification of maize RNA polymerase IV and V subtypes via alternative catalytic subunits. Cell Reports. 2014; 9(1), pp.378–390.

Haberer G, Kamal N, Bauer E, Gundlach H, Fischer I, Seidel MA, et al. European maize genomes highlight intraspecies variation in repeat and gene content. Nat Genet. 2020;52: 950–957.

Hagmann J, Becker C, Muller J, Stegle O, Meyer RC, Wang G, Schneeberger K, Fitz J, Altmann T, Bergelson J, et al.: Century-scale methylome stability in a recently diverged Arabidopsis thaliana lineage. *PLoS Genet* 2015, 11:e1004920.

Hale, C.J. et al. A novel Snf2 protein maintains trans-generational regulatory states established by paramutation in maize. *PLoS biology*. 2007; 5(10), pp.2156–2165.

Han Y., S. Qin, and S. R. Wessler, 2013 Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. BMC Genomics 14: 71.

Han Z, Crisp PA, Stelpflug S, Kaeppler SM, Li Q, Springer NM: Heritable Epigenomic Changes to the Maize Methylome Resulting from Tissue Culture. *Genetics* 2018, doi:10.1534/genetics.118.300987.

Hansey C. N., J. M. Johnson, R. S. Sekhon, S. M. Kaeppler, and N. de Leon, 2011 Genetic Diversity of a Maize Association Population with Restricted Phenology. Crop Sci. 51: 704–715.

Haring, M. et al. The role of DNA methylation, nucleosome occupancy and histone modifications in paramutation. *The Plant Journal: for cell and molecular biology*. 2010; 63(3), pp.366–378.

Haun, W.J. & Springer, N.M. Maternal and paternal alleles exhibit differential histone methylation and acetylation at maize imprinted genes. *The Plant Journal: for cell and molecular biology*. 2008

Haun, W.J. et al. Genomic imprinting, methylation and molecular evolution of maize Enhancer of zeste (Mez) homologs. *The Plant Journal: for cell and molecular biology*. 2007; 49(2), pp.325–337.

Heard, E. & Martienssen, R.A. Transgenerational epigenetic inheritance: myths and mechanisms. *Cell*. 2014; 157(1), pp.95–109.

Henderson, I.R. et al. The De novo cytosine methyltransferase DRM2 requires intact UBA domains and a catalytically mutated paralog DRM3 during RNA-directed DNA methylation in Arabidopsis thaliana. *PLoS Genetics*. 2010; 6(10), pp.1–11.

Hermon, P. et al. Activation of the imprinted Polycomb Group Fie1 gene in maize endosperm requires demethylation of the maternal allele. *Plant Molecular Biology*. 2007; 64(4), pp.387–395.

Hirsch C, Hirsch CD, Brohammer AB, Bowman MJ, Soifer I, Barad O, et al. Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize. Plant Cell. 2016; tpc.00353.2016.

Hirsch C. N., J. M. Foerster, J. M. Johnson, R. S. Sekhon, G. Muttoni, et al., 2014 Insights into the Maize Pan-Genome and Pan-Transcriptome. Plant Cell 26: 121–135.

Hofmeister BT, Denkena J, Colomé-Tatché M, Shahryary Y, Hazarika R, Grimwood J, Mamidi S, Jenkins J, Grabowski PP, Sreedasyam A, et al.: A genome assembly and the somatic genetic and epigenetic mutation rate in a wild long-lived perennial Populus trichocarpa. *Genome Biol* 2020, 21:259.

Hofmeister BT, Lee K, Rohr NA, Hall DW, Schmitz RJ. Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. Genome Biol. Genome Biology; 2017;18: 1–16.

Hollick JB: Paramutation and related phenomena in diverse species. *Nature reviews Genetics* 2017, 18:5–23.

Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression.
Genome Res. 2009;19: 1419–1428.

Hsieh, T.-F. et al. Genome-Wide Demethylation of Arabidopsis Endosperm. *Science*. 2009; 324(5933), pp.1451–1454

Ito H, Kakutani T. Control of transposable elements in Arabidopsis thaliana. Chromosome Res. 2014; 22: 217–223.

Jackson JP, Lindroth AM, Cao X, Jacobsen SE. Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. Nature. 2002; 416: 556–560.

Ji L, Jordan WT, Shi X, Hu L, He C, Schmitz RJ: TET-mediated epimutagenesis of the Arabidopsis thaliana methylome. *Nat Commun* 2018, 9:895.

Ji L, Mathioni SM, Johnson S, Tucker D, Bewick AJ, Do Kim K, Daron J, Keith Slotkin R, Jackson SA, Parrott WA, et al.: Genome-Wide Reinforcement

of DNA Methylation Occurs during Somatic Embryogenesis in Soybean. *The Plant Cell* 2019, 31:2315– 2331.

Jia, Y. et al. Loss of RNA-dependent RNA polymerase 2 (RDR2) function causes widespread and unexpected changes in the expression of transposons, genes, and 24- nt small RNAs. PLoS genetics. 2009; 5(11), p.e1000737.

Jiang C, Mithani A, Belfield EJ, Mott R, Hurst LD, Harberd NP: Environmentally responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and epimutations. *Genome Res* 2014, 24:1821–1829.

Jiang N., and S. R. Wessler, 2001 Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements. Plant Cell 13: 2553–2564.

Jiang N., Z. Bao, X. Zhang, S. R. Eddy, and S. R. Wessler, 2004 Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569–573.

Jiang S-H, Sun Q-G, Chen M, Wang N, Xu H-F, Fang H-C, Wang Y-C, Zhang Z-Y, Chen X-S: Methylome and transcriptome analyses of apple fruit somatic mutations reveal the difference of red phenotype. *BMC Genomics* 2019, 20:117.

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Campbell MS, et al. The complex sequence landscape of maize revealed by single molecule technologies. 2016; 1–19.

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference genome with single-molecule technologies. Nature. 2017;546: 524–527.

Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, Bulski A, Albuisson J, Heredia F, Audigier P, et al.: Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet* 2009, 5:e1000530.

Johnson LM, Du J, Hale CJ, Bischof S, Feng S, Chodavarapu RK, Zhong X, Marson G, Pellegrini M, Segal DJ, et al.: SRA- and SET-domain-containing proteins link RNA polymerase V occupancy to DNA methylation. *Nature* 2014, 507:124–128.

Jordan I. K., I. King Jordan, I. B. Rogozin, G. V. Glazko, and E. V. Koonin, 2003 Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends in Genetics 19: 68–72.

Jullien, P.E. et al. DNA methylation dynamics during sexual reproduction in Arabidopsis thaliana. *Current biology: CB*. 2012; 22(19), pp.1825–1830.

Kaeppler, S.M. & Phillips, R.L. Tissue culture-induced DNA methylation variation in maize.

Proc. Natl. Acad. Sci. U. S. A. 1993; 90(19), pp.8773–8776.

Kaeppler, S.M., Kaeppler, H.F. & Rhee, Y. Epigenetic aspects of somaclonal variation in plants. *Plant Molecular Biology*. 2000; 43(2–3), pp.179–188.

Kawakatsu T, Huang SS, Jupe F, Sasaki E, Schmitz RJ, Urich MA, Castanon R, Nery JR, Barragan C, He Y, et al.: Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. *Cell* 2016, 166:492–505.

Kawakatsu, T. et al. Unique cell-type-specific patterns of DNA methylation in the root meristem. *Nature Plants*. 2016; 2, p.16058.

Kermicle, J.L. Dependence of the R-Mottled Aleurone Phenotype in Maize on Mode of Sexual Transmission. *Genetics*. 1970; 66(1), pp.69–85.

Kim MY, Zilberman D. DNA methylation as a system of plant genomic immunity. Trends Plant Sci. 2014; 19: 320–326.

Kolkman JM, Conrad LJ, Farmer PR, et al. Distribution of Activator (Ac) throughout the maize genome for use in regional mutagenesis. Genetics. 2005;169(2):981-995. doi:10.1534/genetics.104.033738

Kooke R, Johannes F, Wardenaar R, Becker F, Etcheverry M, Colot V, Vreugdenhil D, Keurentjes JJ: Epigenetic basis of morphological variation and phenotypic plasticity in Arabidopsis thaliana. *Plant Cell* 2015, 27:337–348.

Kremling K. A. G., S.-Y. Chen, M.-H. Su, N. K. Lepak, M. C. Romay, et al., 2018 Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature 555: 520–523.

Lämke J, Bäurle I: Epigenetic and chromatin-based mechanisms in environmental stress adaptation and stress memory in plants. *Genome Biol* 2017, 18:124.

Langmead B. Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics. 2010; Chapter 11: Unit 11.7.

Latutrie M, Gourcilleau D, Pujol B: Epigenetic variation for agronomic improvement: an opportunity for vegetatively propagated crops. *Am J Bot* 2019, 106:1281–1284.

Lauria, M. et al. Extensive maternal DNA hypomethylation in the endosperm of Zea mays.
The Plant Cell. 2004; 16(2), pp.510–522.

Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature reviews Genetics. 2010;11: 204–220.

Law, J.A. et al. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. Nature. 2013; 498(7454), pp.385–389.

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34: 3094–3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. Welcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA.; 2009; 25: 2078–2079.

Li H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, et al., 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Li Q, Eichten SR, Hermanson PJ, Springer NM: Inheritance patterns and stability of DNA methylation variation in maize near-isogenic lines. *Genetics* 2014b, 196:667–676.

Li Q, Eichten SR, Hermanson PJ, Zaunbrecher VM, Song J, Wendt J, et al. Genetic perturbation of the maize methylome. Plant Cell. 2014a; 26: 4602–4616.

Li Q, Gent JI, Zynda G, Song J, Makarevitch I, Hirsch CD, et al. RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. Proc Natl Acad Sci U S A. 2015a;112: 14728–14733.

Li Q. et al. Examining the Causes and Consequences of Context-Specific Differential DNA Methylation in Maize. *Plant Physiology*. 2015b; 168(4), pp.1262–1274.

Li Q. et al. Post-conversion targeted capture of modified cytosines in mammalian and plant genomes. *Nucleic Acids Research*. 2015c; 43(12), pp.1–16.

Li W-F, Ning G-X, Mao J, Guo Z-G, Zhou Q, Chen B-H: Whole-genome DNA methylation patterns and complex associations with gene expression associated with anthocyanin biosynthesis in apple fruit skin. *Planta* 2019, 250:1833–1847.

Lisch D, Bennetzen JL. Transposable element origins of epigenetic gene regulation. Curr Opin Plant Biol. 2011;14: 156–161.

Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2013;14: 49– 61.

Lisch D. Mutator and MULE Transposons. Microbiol Spectr. 2015;3(2): MDNA3-2014. doi: 10.1128/microbiolspec.MDNA3-0032-2014

Lisch, D. et al. A mutation that prevents paramutation in maize also reverses Mutator transposon methylation and silencing. *Proceedings of the National Academy of Sciences of the United States of America*. 2002; 99(9), pp.6130–6135.

Lister, R. et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*. 2008; 133(3), pp.523–536.

Liu S, Yeh CT, Ji T, et al. Mu transposon insertion sites and meiotic recombination events co- localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet. 2009;5(11): e1000733. doi: 10.1371/journal.pgen.1000733

Liu, R. et al. A DEMETER-like DNA demethylase governs tomato fruit ripening. *Proceedings of the National Academy of Sciences of the United States of America*. 2015; 112(34), pp.10804–10809.

Loenen, W.A.M. & Raleigh, E.A. The other face of restriction: Modification-dependent enzymes. Nucleic Acids Research. 2014; 42(1), pp.56–69.

Love M, Anders S, Huber W. Differential analysis of count data--the DESeq2 package. Genome Biol. 2014;15: 10–1186.

Lowe CB, Bejerano G, Haussler D (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci USA 104: 8005–8010

Lu W, Xiao L, Quan M, Wang Q, El-Kassaby YA, Du Q, Zhang D: Linkage-linkage disequilibrium dissection of the epigenetic quantitative trait loci (epiQTLs) underlying growth and wood properties in Populus. *New Phytologist* 2020, 225:1218–1233.

Lu Z., A. P. Marand, W. A. Ricci, C. L. Ethridge, X. Zhang, et al., 2019 The prevalence, evolution and chromatin signatures of plant regulatory elements. Nat Plants 5: 1250– 1259.

Lu, Y., Rong, T. & Cao, M. Analysis of DNA methylation in different maize tissues. *Journal of Genetics and Genomics*. 2008; 35(1), pp.41–48.

Maher B: Personal genomes: The case of the missing heritability. *Nature* 2008, 456:18–21.

Makarevitch I., A. J. Waters, P. T. West, M. Stitzer, C. N. Hirsch, et al., 2015 Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS Genet. 11: e1004915.

Makarevitch, I. et al. Genomic distribution of maize facultative heterochromatin marked by trimethylation of H3K27. *The Plant Cell*. 2013; 25(3), pp.780–793.

Makarevitch, I. et al. Natural variation for alleles under epigenetic control by the maize chromomethylase Zmet2. *Genetics*. 2007; 177(2), pp.749–760.

Mao H., H. Wang, S. Liu, Z. Li, X. Yang, et al., 2015 A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. Nat. Commun. 6: 8326.

Marand AP, Chen Z, Gallavotti A, Schmitz RJ. A cis-regulatory atlas in maize at single-cell resolution. 2020. doi:10.1101/2020.09.27.315499

Martienssen RA, Colot V. DNA methylation and epigenetic inheritance in plants and filamentous fungi. Science. 2001;293: 1070–1074.

Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, et al. A transposon- induced epigenetic change leads to sex determination in melon. Nature. 2009;461: 1135–1138.

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.

EMBnet.journal. 2011;17: 10–12.

Maside X., C. Bartolomé, and B. Charlesworth, 2002 S-element insertions are associated with the evolution of the Hsp70 genes in Drosophila melanogaster. Curr. Biol. 12: 1686– 1691.

Matzke, M.A. & Mosher, R.A. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nature*, 2014; 15(June).

Mazaheri M., M. Heckwolf, B. Vaillancourt, J. L. Gage, B. Burdo, et al., 2019 Genome-wide association analysis of stalk biomass and anatomical traits in maize. BMC Plant Biol. 19: 45.

McCarty DR, Latshaw S, Wu S, Suzuki M, Hunter CT, Avigne WT, et al. Mu-seq: sequence-based mapping and identification of transposon induced mutations. PLoS One. 2013;8: e77172.

McClintock B., Chromosome organization and genic expression. Cold Spring Harb. Symp.

Quant. Biol. 1951; 16: 13–47.

McClintock, B. Aspects of gene regulation in maize. *Carnegie Inst Wash Year Book*. 1964; 63, pp.592–602.

MCCLINTOCK, B. Controlling elements and the gene. *Cold Spring Harbor symposia on quantitative biology*. 1956; 21, pp.197–216.

McCue, A.D. et al. ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the establishment of DNA methylation. The EMBO Journal. 2014; 34(1), pp.20– 35.

Mei, W. et al. A Comprehensive Analysis of Alternative Splicing in Paleopolyploid Maize.

Frontiers in Plant Science. 2017; 8(May), pp.1–19.

Melquist S, Bender J: An internal rearrangement in an Arabidopsis inverted repeat locus impairs DNA methylation triggered by the locus. *Genetics* 2004, 166:437–448.

Melquist S, Luff B, Bender J: Arabidopsis PAI gene arrangements, cytosine methylation and expression. *Genetics* 1999, 153:401–413.

Michael T. P., and S. Jackson, 2013 The first 50 plant genomes. Plant Genome 6.

Morgan HD, Sutherland HG, Martin DI, Whitelaw E. Epigenetic inheritance at the agouti locus in the mouse. Nat Genet. 1999;23: 314–318.

Naito K., F. Zhang, T. Tsukiyama, H. Saito, C. N. Hancock, et al., 2009 Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature 461: 1130–1134.

Nguyen HM, Kim M, Ralph PJ, Marín-Guirao L, Pernice M, Procaccini G: Stress Memory in

Seagrasses: First Insight Into the Effects of Thermal Priming and the Role of Epigenetic Modifications. *Frontiers in Plant Science* 2020, 11.

Niederhuth C. E., A. J. Bewick, L. Ji, M. S. Alabady, K. D. Kim, et al., 2016 Widespread natural variation of DNA methylation within angiosperms. bioRxiv http://dx.

Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Page JT, et al. Widespread natural variation of DNA methylation within angiosperms. Genome Biol. 2016; http://dx.doi.org/10.1101/045880: 194.

Nishihara H, Smit AFA, Okada N (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res 16: 864–874

Noshay JM, Anderson SN, Zhou P, Ji L, Ricci W, Lu Z, Stitzer MC, Crisp PA, Hirsch CN, Zhang X, et al.: Monitoring the interplay between transposable element families and DNA methylation in maize. *PLoS Genet* 2019, 15:e1008291.

Noshay JM, Crisp PA, Springer NM. The Maize Methylome. In: Bennetzen J, Flint-Garcia S, Hirsch C, Tuberosa R, editors. The Maize Genome. Cham: Springer International Publishing; 2018. pp. 81–96.

O'Connor C. H., Y. Qiu, R. D. Coletta, P. J. Monnahan, et al., Population Level Variation of Transposable Elements in a Maize Diversity Panel. BioRxiv. 2020; 10.1101/2020.09.25.314401

Ocaña J, Walter B, Schellenbaum P: Stable MSAP markers for the distinction of Vitis vinifera cv Pinot noir clones. *Mol Biotechnol* 2013, 55:236–248.

Oka R, Zicola J, Weber B, Anderson SN, Hodgman C, Gent JI, et al. Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. Genome Biol. 2017;18: 137.

Ong-Abdullah M, Ordway JM, Jiang N, Ooi SE, Kok SY, Sarpan N, Azimi N, Hashim AT, Ishak Z, Rosli SK, et al.: Loss of Karma transposon methylation

underlies the mantled somaclonal variant of oil palm. *Nature* 2015, 525:533–537.

Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M: The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science* 2010, 327:92–94.

Palmer, L.E. Maize Genome Sequencing by Methylation Filtration. *Science*. 2003; 302(5653), pp.2115–2117.

Panda K, Ji L, Neumann DA, Daron J, Schmitz RJ, Slotkin RK. Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA methylation. Genome Biol. 2016;17: 170–016–1032–y.

Panda, K. & Slotkin, R.K. Proposed mechanism for the initiation of transposable element silencing by the RDR6-directed DNA methylation pathway. *Plant signaling & behavior*. 2013; 8(8), pp.8–10.

Papa, C.M. Maize Chromomethylase Zea methyltransferase2 Is Required for CpNpG Methylation. *the Plant Cell Online*. 2001; 13(8), pp.1919–1928.

Papikian A, Liu W, Gallego-Bartolomé J, Jacobsen SE: Site-specific manipulation of Arabidopsis loci using CRISPR-Cas9 SunTag systems. *Nature Communications* 2019, 10.

Park, M., Keung, A.J. & Khalil, A.S. The epigenome: the next substrate for engineering. *Genome biology*. 2016; 17(1), pp.183–185.

Pecinka, A. & Scheid, O.M. Stress-Induced Chromatin Changes: A Critical View on their Heritability. *Plant and Cell Physiology*. 2012; p.pcs044.

Phillips, R.L., Kaeppler, S.M. & Olhoft, P. (1994) Genetic instability of plant tissue cultures: breakdown of normal controls. *Proceedings of the National Academy of Sciences of the United States of America*, 91(12), pp.5222–5226.

Picard Tools - By Broad Institute. [cited 7 Mar 2019]. Available: http://broadinstitute.github.io/picard/

Piffanelli P, Droc G, Mieulet D, Lanau N, Bès M, Bourgeois E, et al. Large-scale characterization of Tos17 insertion sites in a rice T-DNA mutant library. Plant Mol Biol. 2007;65: 587–601.

Plongthongkum N, Diep DH, Zhang K: Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat Rev Genet* 2014, 15:647–661.

Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA, et al. The Arabidopsis thaliana mobilome and its impact at the species level. Elife. 2016;5. doi:10.7554/eLife.15716

Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, Stein L, et al. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nat Genet. 1999;23: 305–308.

Rabinowicz, P.D. et al. Differential methylation of genes and repeats in land plants. *Genome Research*. 2005; 15(10), pp.1431–1440.

Regulski M, Lu Z, Kendall J, Donoghue MT, Reinders J, Llaca V, Deschamps S, Smith A, Levy D, McCombie WR, et al.: The maize methylome influences

mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Res* 2013, 23:1651–1662.

Reinders J, Wulff BB, Mirouze M, Mari-Ordonez A, Dapp M, Rozhon W, Bucher E, Theiler G, Paszkowski J: Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes. *Genes Dev* 2009, 23:939–950.

Ricci W. A., Z. Lu, L. Ji, A. P. Marand, C. L. Ethridge, et al., Widespread long-range cis- regulatory elements in the maize genome. Nat Plants. 2019; 5: 1237–1249.

Richards EJ: Inherited epigenetic variation--revisiting soft inheritance. *NatRevGenet* 2006, 7:395–401.

Rodgers-Melnick E., D. L. Vera, H. W. Bass, and E. S. Buckler, Open chromatin reveals the functional maize genome. Proc. Natl. Acad. Sci. U. S. A. 2016; 113: E3177–84.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. Nat Genet. 1998;20: 43–45.

Sarpan N, Taranenko E, Ooi S-E, Low E-TL, Espinoza A, Tatarinova TV, Ong-Abdullah M: DNA methylation changes in clonally propagated oil palm. *Plant Cell Rep* 2020, 39:1219–1233.

Schmid-Siegert E, Sarkar N, Iseli C, Calderon S, Gouhier-Darimont C, Chrast J, Cattaneo P, Schütz F, Farinelli L, Pagni M, et al.: Low number of fixed somatic mutations in a long-lived oak tree. *Nat Plants* 2017, 3:926–929.

Schmitz RJ, He Y, Valdes-Lopez O, Khan SM, Joshi T, Urich MA, Nery JR, Diers B, Xu D, Stacey G, et al.: Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Res* 2013, 23:1663–1674.

Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, Schork NJ, Ecker JR: Transgenerational epigenetic instability is a source of novel methylation variants. *Science* 2011, 334:369–373.

Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ, et al.: Patterns of population epigenomic diversity. *Nature* 2013, 495:193–198.

Schmitz, R.J. & Ecker, J.R. Epigenetic and epigenomic variation in Arabidopsis thaliana.
Trends in Plant Science. 2012; 17, pp.149–154.

Schnable P. S., D. Ware, R. S. Fulton, J. C. Stein, S. Pasternak, et al., 2009 The B73 Maize Genome: Complexity, Diversity, and Dynamics. 326.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. Center for Plant Genomics, Iowa State University, Ames, IA 50011, USA.; 2009;326: 1112–1115.

Schönberger B, Chen X, Mager S, Ludewig U: Site-Dependent Differences in DNA Methylation and Their Impact on Plant Establishment and Phosphorus Nutrition in Populus trichocarpa. *PLoS One* 2016, 11:e0168623.

Secco D, Wang C, Shou H, Schultz MD, Chiarenza S, Nussaume L, Ecker JR, Whelan J, Lister R: Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements. *Elife* 2015, 4:10.7554/eLife.09343.

Selinger, D.A. & Chandler, V.L. B-Bolivia, an allele of the maize b1 gene with variable expression, contains a high copy retrotransposon-related sequence immediately upstream. *Plant Physiology*. 2001; 125(3), pp.1363–1379.

Shahryary Y, Symeonidi A, Hazarika RR, Denkena J: AlphaBeta: Computational inference of epimutation rates and spectra from high-throughput DNA methylation data in plants. *bioRxiv*. 2020

Sheffield N. C., R. E. Thurman, L. Song, A. Safi, J. A. Stamatoyannopoulos, et al., Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Res. 2013; 23: 777–788.

Shen Y, Zhang J, Liu Y, Liu S, Liu Z, Duan Z, Wang Z, Zhu B, Guo Y-L, Tian Z: DNA methylation footprints during soybean domestication and improvement. *Genome Biol* 2018, 19:128.

Sidorenko, L. V & Peterson, T. Transgene-induced silencing identifies sequences involved in the establishment of paramutation of the maize p1 gene. *The Plant Cell*. 2001; 13(2), pp.319–335.

Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nature reviews Genetics. 2007;8: 272–285.

Smith, A.M., Hansey, C.N. & Kaeppler, S.M. TCUP: A Novel hAT Transposon Active in Maize Tissue Culture. *Frontiers in plant science*. 2012; 3, p.6.

Springer N. M., D. Lisch, and Q. Li, 2016 Creating Order from Chaos: Epigenome Dynamics in Plants with Complex Genomes. Plant Cell 28: 314–325.

Springer N. M., S. N. Anderson, C. M. Andorf, K. R. Ahern, F. Bai, et al., 2018 The maize W22 genome provides a foundation for functional genomics and transposon biology. Nat. Genet. https://doi.org/10.1038/s41588-018-0158-0

Springer NM, Schmitz RJ. Exploiting induced and natural epigenetic variation for crop improvement. Nat Rev Genet. 2017; doi:10.1038/nrg.2017.45

Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y, et al. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. 2009;5: e1000734.

Stam, M. Paramutation: A Heritable Change in Gene Expression by Allelic Interactions in Trans. *Molecular Plant*. 2009; 2(4), pp.578–588.

Stelpflug, S.C. et al. Consistent and heritable alterations of DNA methylation are induced by tissue culture in maize. *Genetics*. 2014; 198(1), pp.209–218.

Stitzer MC, Anderson SN, Springer NM, Ross-Ibarra J. The Genomic Ecosystem of Transposable Elements in Maize. bioRxiv. 2019. doi:10.1101/559922

Stonaker, J.L. et al. Diversity of Pol IV function is defined by mutations at the maize rmr7 locus. *PLoS Genetics*. 2009; 5(11).

Stroud H, Ding B, Simon SA, Feng S, Bellizzi M, Pellegrini M, Wang GL, Meyers BC, Jacobsen SE: Plants regenerated from tissue culture contain stable epigenome changes in rice. *Elife* 2013, 2:e00354.

Stroud, H. et al. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. *Cell*, 2013;152(1–2), pp.352–364.

Stroud, H. et al. Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. *Nature structural & molecular biology*. 2014; 21(1), pp.64–72.

Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, Lister R. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. Elife. 2016;5: 10.7554/eLife.20777.

Studer A., Q. Zhao, J. Ross-Ibarra, and J. Doebley, 2011 Identification of a functional transposon insertion in the maize domestication gene tb1. Nat. Genet. 43: 1160– 1163.

Sultana T, Zamborlini A, Cristofari G, Lesage P. Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet. 2017;18: 292–308.

Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, et al. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat Genet. 2018;50: 1289–1295.

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, et al. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res. 2010;20: 1689–1699.

Taudt A, Colome-Tatche M, Johannes F: Genetic sources of population epigenomic variation. *Nature reviews Genetics* 2016, 17:319–332.

The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408: 796–815.

To, T.K., Saze, H. & Kakutani, T. DNA methylation within transcribed regions. *Plant Physiology*. 2015; 4, p.pp.00543.2015.

Tsukahara S, Kawabe A, Kobayashi A, Ito T, Aizu T, Shin-i T, et al. Centromere-targeted de novo integrations of an LTR retrotransposon of Arabidopsis lyrata. Genes Dev. 2012;26: 705–713.

Underwood, C.J., Henderson, I.R. & Martienssen, R.A. Genetic and epigenetic variation of transposable elements in Arabidopsis. *Current Opinion in Plant Biology*. 2017; 36, pp.135–141.

van der Graaf, A. et al. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proceedings of the National Academy of Sciences of the United States of America*. 2015; 112(21), pp.6676–6681.

Vaughn MW, Tanurd Ic M, Lippman Z, Jiang H, Carrasquillo R, Rabinowicz PD, Dedhia N, McCombie WR, Agier N, Bulski A, et al.: Epigenetic Natural Variation in Arabidopsis thaliana. *PLoS Biol* 2007, 5:e174.

Vollbrecht E, Duvick J, Schares JP, Ahern KR, Deewatthanawong P, Xu L, et al. Genome- wide distribution of transposed Dissociation elements in maize. Plant Cell. 2010;22: 1667–1685.

Walker, E.L. Paramutation of the r1 locus of maize is associated with increased cytosine methylation. *Genetics*. 1998; 148(4), pp.1973–1981.

Walley JW, Sartor RC, Shen Z, Schmitz RJ, Ecker JR, Briggs SP. Integration of omic networks in a developmental atlas of maize. Science. 2016;353: 814–818.

Wang Q., and H. K. Dooner, 2006 Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. Proc. Natl. Acad. Sci. U. S. A. 103: 17644– 17649.

Wang, P. et al. Genome-wide high-resolution mapping of DNA methylation identifies epigenetic variation across embryo and endosperm in Maize (Zea may). *BMC Genomics*. 2015; 16, p.21.

Waters, A.J. et al. Parent-of-Origin Effects on Gene Expression and DNA Methylation in the Maize Endosperm. *The Plant Cell*. 2011; 23(12), pp.4221–4233.

Weil C., and R. Martienssen, 2008 Epigenetic interactions between transposons and genes: lessons from plants. Curr. Opin. Genet. Dev. 18: 188–192.

West PT, Li Q, Ji L, Eichten SR, Song J, Vaughn MW, et al. Genomic distribution of H3K9me2 and DNA methylation in a maize genome. PLoS One. 2014;9: e105267.

Wibowo A, Becker C, Marconi G, Durr J, Price J, Hagmann J, Papareddy R, Putra H, Kageyama J, Becker J, et al.: Hyperosmotic stress memory in Arabidopsis is mediated by distinct epigenetically labile sites in the genome and is restricted in the male germline by DNA glycosylase activity. *Elife* 2016, 5.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nature reviews Genetics. 2007;8: 973–982.

Wittmeyer K, Cui J, Chatterjee D, Lee T-F, Tan Q, Xue W, et al. The Dominant and Poorly Penetrant Phenotypes of Maize Unstable factor for orange1 Are Caused by DNA Methylation Changes at a Linked Transposon. Plant Cell. 2018;30: 3006–3023.

Wolff, P. et al. High-Resolution Analysis of Parent-of-Origin Allelic Expression in the Arabidopsis Endosperm. *PLoS genetics*. 2011; 7(6), p.e1002126.

Woo, H.R., Dittmer, T.A. & Richards, E.J. Three SRA-domain methylcytosine-binding proteins cooperate to maintain global CpG methylation and epigenetic silencing in Arabidopsis. *PLoS Genetics*. 2008; 4(8).

Woodhouse, M.R., Freeling, M. & Lisch, D. Initiation, establishment, and maintenance of heritable MuDR transposon silencing in maize are mediated by distinct factors.
PLoS biology. 2006a; 4(10), p.e339.

Woodhouse, M.R., Freeling, M. & Lisch, D. The mop1 (mediator of paramutation1) mutant progressively reactivates one of the two genes encoded by the MuDR transposon in maize. *Genetics*. 2006b; 172(1), pp.579–592.

Wyler M, Stritt C, Walser JC, et al. Impact of transposable elements on methylation and gene expression across natural accessions of Brachypodium distachyon. bioRxiv 2020.06.16.154047; doi: 10.1101/2020.06.16.154047 Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009;10: 232.

Xie H, Konate M, Sai N, Tesfamicael KG, Cavagnaro T, Gilliham M, Breen J, Metcalfe A, Stephen JR, De Bei R, et al.: Global DNA Methylation Patterns Can Play a Role in Defining Terroir in Grapevine (Vitis vinifera cv. Shiraz). *Front Plant Sci* 2017, 8:1860.

Xu G, Lyu J, Li Q, Liu H, Wang D, Zhang M, Springer NM, Ross-Ibarra J, Yang J: Evolutionary and functional genomics of DNA methylation in maize domestication and improvement. *Nat Commun* 2020, 11:5539.

Xu J, Chen G, Hermanson PJ, Xu Q, Sun C, Chen W, Kan Q, Li M, Crisp PA, Yan J, et al.: Population-level analysis reveals the widespread occurrence and phenotypic consequence of DNA methylation variation not tagged by genetic variation in maize. *Genome Biol* 2019, 20:243.

Yang Q., Z. Li, W. Li, L. Ku, C. Wang, et al., 2013 CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the post domestication spread of maize. Proc. Natl. Acad. Sci. U. S. A. 110: 16969–16974.

Yoder JA, Walsh CP, Bestor TH. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. 1997;13: 335–340.

Yuan Y, SanMiguel PJ, Bennetzen JL. Methylation-spanning linker libraries link gene-rich regions and identify epigenetic boundaries in Zea mays. Genome Res. 2002;12: 1345–1349.

Zemach, A. et al. Local DNA hypomethylation activates genes in rice endosperm. *Proceedings of the National Academy of Sciences of the United States of America*. 2010; 107(43), pp.18729–18734.

Zemach, A. et al. The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell*. 2013; 153(1), pp.193–205.

Zerjal T., A. Rousselet, C. Mhiri, V. Combes, D. Madur, et al., 2012 Maize genetic diversity and association mapping using transposable element insertion polymorphisms. Theor. Appl. Genet. 124: 1521–1537.

Zhang H, Lang Z, Zhu J-K. Dynamics and function of DNA methylation in plants. Nat Rev Mol Cell Biol. 2018;19: 489–506.

Zhang P., W. B. Allen, N. Nagasawa, A. S. Ching, E. P. Heppard, et al., 2012 A transposable element insertion within ZmGE2 gene is associated with increase in embryo to endosperm ratio in maize. Theor. Appl. Genet. 125: 1463–1471.

Zhang X, Clarenz O, Cokus S, Bernatavichute YV, Pellegrini M, Goodrich J, et al. Whole- genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. PLoS Biol. 2007;5: e129.

Zhang Y-Y, Latzel V, Fischer M, Bossdorf O: Understanding the evolutionary potential of epigenetic variation: a comparison of heritable phenotypic variation in epiRILs, RILs, and natural ecotypes of Arabidopsis thaliana. *Heredity* 2018, 121:257–265.

Zhang YY, Fischer M, Colot V, Bossdorf O: Epigenetic variation creates potential for evolution of plant phenotypic plasticity. *New Phytol* 2013, 197:314–322.

Zhang, H. & Zhu, J.K. Active DNA demethylation in plants and animals. *Cold Spring Harbor symposia on quantitative biology*. 2012; 77, pp.161–173.

Zhang, M. et al. Genome-wide high-resolution parental-specific DNA and histone methylation maps uncover patterns of imprinting regulation in maize. *Genome research*. 2014; 24(1), pp.167–176.

Zhang, M. et al. Tissue-specific differences in cytosine methylation and their association with differential gene expression in sorghum. *Plant Physiology*. 2011; 156(4), pp.1955–1966.

Zhao H, Zhang W, Chen L, et al. Proliferation of Regulatory DNA Elements Derived from Transposable Elements in the Maize Genome. Plant Physiol. 2018;176(4):2789- 2803. doi:10.1104/pp.17.01467

Zhao L, Xie L, Zhang Q, Ouyang W, Deng L, Guan P, Ma M, Li Y, Zhang Y, Xiao Q, et al.: Integrative analysis of reference epigenomes in 20 rice varieties. *Nat Commun* 2020, 11:2658.

Zheng X, Chen L, Xia H, Wei H, Lou Q, Li M, Li T, Luo L: Transgenerational epimutations induced by multi-generation drought imposition mediate rice plant's adaptation to drought condition. *Sci Rep* 2017, 7:39843.

Zilberman, D. et al. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature genetics*. 2007; 39(1), pp.61–69.

Zong, W. et al. Genome-wide profiling of histone H3K4-tri-methylation and gene expression in rice under drought stress. *Plant Molecular Biology*. 2013; 81(1–2), pp.175–188.